The Dissertation Committee for Austin Garig Meyer
certifies that this is the approved version of the following dissertation:

The Role of Structure in Protein Evolution

Committee:

Claus O. Wilke, Supervisor

Karen Browning

Andrew Ellington

David Hoffman

Pengyu Ren

The Role of Structure in Protein Evolution

by

Austin Garig Meyer, B.A., B.S., M.S.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

December 2014

To my family and friends who have put up with my extended time in school.

Acknowledgments

First, I would like to acknowledge my advisor Claus Wilke. He did a great job of initially offering project ideas, pushing harder when they weren't going well, supporting my ideas both financially and intellectually, helping fix issues when projects were failing, and doing everything to make me familiar with the job of research. More importantly, I cannot imagine many advisors sticking with a student through the first two years of medical school. No student could expect anything more.

In addition, I would like to acknowledge my partner Brittany Rosales. She has been with me through the years of undergraduate, master's, and doctoral programs. She has put up with my almost annual move from Austin to Lubbock and back. Moreover, she has read every manuscript I have produced since my second year of undergraduate school contributing ideas and edits as needed. It would not have been possible to complete this without her.

I would also like to acknowledge Andy Ellington. Despite never officially working in his lab, Andy has intellectually and financially supported my projects from the beginning of my time at UT. Bryan Sutton helped me continue my research while in medical school by giving me access to computer resources and physical lab space. Also, as he was my advisor for the M.S. degree, he started my interest in computational and structural biology. Several

of my fellow Wilke lab comrades, in particular Matthew Tien, have been critical to advancing my training, making me think through projects, and giving me the opportunity to deepen my understanding by training others.

The Role of Structure in Protein Evolution

Austin Garig Meyer, Ph.D.
The University of Texas at Austin, 2014

Supervisor: Claus O. Wilke

Identifying sites under evolutionary pressure and predicting the effects of substitutions at those sites are among the greatest standing problems in bioinformatics and computational biology. Moreover, the two problems have traditionally been separated by the enormous chasm that exists between molecular evolutionary biologists interested in the evolutionary process and theoretical chemists interested in free energy changes. As a result, identifying sites under selective pressure has most often left out any semblance of structural biology and biochemistry; likewise, theoretical chemistry tends to rely strictly on first principles calculations rather than thinking first about biologically simple and interpretable results. Here, I have tried to integrate these two intuitions with regard to protein function and evolution. First, I developed a model that implements structural measurements into a traditional structure-blind molecular evolutionary model. This structure-aware model performs significantly better at identifying sites under both purifying and diversifying selection than

its structure-blind counter part. Second, I go further to understand the extent to which structural features of any kind can predict the evolutionary process. By comparing site-wise evolution between human and avian influenza, I find that structural features can account for 24% to 36% of the evolutionary pressure on influenza hemagglutinin. Third, I developed a computational method based on first principles molecular dynamics simulations to predict the biological effect of substitutions in the Machupo virus–Human receptor protein–protein interface. I found that relatively simple energetic proxies offer a reasonable substitute for rigorous free energy calculations; such simple proxies could allow non-experts to naively implement first principles methods without being forced to consider all possible degrees of freedom for *post hoc* calculations.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Nearly since the discovery of the function of DNA and the subsequent push to decode the information it contained, there have been attempts to understand the evolutionary process through sequence analysis. For the flu, starting from a very naive place, some early efforts largely focused on speculation about the likely functional result of amino acid substitutions. One such attempt analyzed the biochemical effect (more hydrophobic to less hydrophobic, etc.) of substitutions without regard for their position in the protein structure [58]. By contrast, several studies focused on the substitution within the context of the final folded protein. For example, one study compared a single sequence of hemagglutinin from the 1968 emergent strain of H3N2 influenza to a small number of subsequent strains, and contrasted amino acid substitutions that carried the biggest putative biochemical changes (charged to uncharged, etc.) [105]. Although this method did pay some dividends in terms of understanding influenza, it provided little information about the evolutionary process of the evolving hemagglutinin protein responding to host pressure. Moreover, each of these studies was almost completely unsystematic and very speculative due to technical and methodological limitations.

More recently, as there has been enormous growth in the number of available genetic sequences, it has become possible to calculate the degree to which sites in individual proteins are under evolutionary pressure [107]. Although perhaps not strictly a method from molecular evolution, one common approach is to use the Shannon information entropy at each site (S_i) defined as,

$$S_i = -\sum_i P_{ij} \ln P_{ij}, \tag{1.1}$$

where j represents each of the 20 possible amino acids and P_{ij} is the frequency of each amino acid j at position i in the alignment. Calculating site-wise entropies for every position in an amino acid alignment has several advantages over speculating about possible functional changes. First, even without an available protein structure, amino acid substitutions are carried out in the context of protein structure, and they implicitly retain a significant amount of structural information in one dimensional sequences. As a result, with only an amino acid alignment, the Shannon entropy can be applied over a broad range of circumstances [93]. Second, the quantity of sequence entropy is relatively easy to understand; a higher number means there is more entropy at a site and thus there is probably a greater amount of diversifying selection.

On the other hand, sequence entropy calculations have several critical problems for application in molecular evolution. Simple information entropy leaves out any notion of an expected difference between two different sites in a protein; it treats every site with the same distribution of amino acids and leaves no room for alternative assumptions. If as a result of the underlying

mutational process, two sites in the same protein have different chances of acquiring substitutions, that difference is not controlled for within the context of sequence entropy. In addition, any concept of evolutionary time and thus the ability to appropriately compare sites with substitutions at different branch lengths is difficult or impossible.

Probably the most important advancement in the field of molecular evolution was the ability to identify sites under positive selection while controlling for the underlying mutational process [107]. Because synonymous changes do not change the amino acid sequence, the chance that they will have a dramatic effect on fitness is very low [107]. Thus, one can calculate the rate at which synonymous substitutions accumulate (dS) at each site in a codon-based alignment. Then, the non-synonymous substitution rate (dN) can be separately calculated for each site and divided by the synonymous rate to determine the substitution excess of non-synonymous changes at each site; that number is referred to as the evolutionary rate $\omega = dN/dS$. Since any site with $\omega > 1$ has an excess of non-synonymous substitutions relative to synonymous substitutions those sites are undergoing positive selection. Furthermore, the quantity ω is relatively easy to integrate into a codon model that includes terms to account for other important aspects of the substitution process. For example, parameters for evolutionary time and the difference between transition and transversion mutations are both included within the model and can be estimated simultaneously via maximum likelihood estimation.

Despite the advancement of molecular evolutionary techniques, one

thing has remained lost from the initial analysis of functional biochemical differences [66, 89]. That is, proteins are three dimensional objects with a large amount of internal flexibility, and they exist within a heterogeneous biochemical milieu. They must fold appropriately to carry out any physiological function at all and they must interface with sometimes dozens of other proteins to carry on a greater functional role within a cell. As a result, an approach based strictly on sequence analysis will always fail to capture at least some aspect of the evolutionary process.

We set out to re-integrate structure into molecular evolution, evaluate the extent to which any structure can constrain evolution, and directly and systematically probe the effects of substitutions in proteins without the wild speculation of earlier analyses. To that end, in chapter 2 we develop a molecular evolutionary model in the image of the previous state of the art. In lieu of the standard ω calculation, we integrate a metric of three dimensional structure directly into the model [66]. We find that our new method performs significantly better than the previous state-of-the-art method. Furthermore, the new model helps us better control for evolutionary heterogeneity of sites caused by biophysical differences in a protein site's local environment [66]. In chapter 3 we more broadly apply our new model to an important test case in influenza hemagglutinin. We find that even in a viral system with an enormous amount of positive evolutionary pressure, structure can account for 24% to 36% of evolutionary rate differences among sites in influenza hemagglutinin [64]. In addition, this number very likely represents the lower limit of

structural constraints on protein evolution. A human protein with its relatively low level of positive selection would be expected to have an enormous amount of evolutionary rate variation explained by structure alone. Moreover, we find that our structural metric can only account for approximately 10% of the evolutionary rate variation in hemagglutinin. Thus, more than two-thirds of the rate variation explained by structure has some other determinant. In chapter 4, starting with previously identified sites in a viral protein–protein interface, we apply steered molecular dynamics simulations to compute the energetic effect of substitutions in the interface between a viral protein and its host receptor [65]. We find that out of 10 mutants tested, 40% could be differentiated from the wild type complex based only on maximum force calculations. In addition, we find a very strong agreement between our method and a much more complicated previously established technique.

This research represents a significant advancement to our earlier understanding of molecular evolution. We have begun the process of bringing molecules back in to the study of molecular evolution. Our model allows us to control for purifying pressure in a way that was not previously possible; our hope is that further work will uncover the remaining structural determinants of evolutionary rate variations. A forthcoming paper by Shahmoradi et al. in the *Journal of Molecular Evolution* will directly address this question, but will not itself reveal the final keystone in the search for constraints on the process of evolution in proteins. In addition, our work using molecular dynamics to probe the effects of mutations in protein–protein interactions provides a man-

ner for novice computational biologists to use advanced *ab initio* techniques with relatively little understanding of the underlying mechanics of free energy methods. Though our initial foray took a large amount of computational power, further refinement of the steered molecular dynamics approach could easily reduce the required cost by an order of magnitude or more. Moreover, such a reduction in cost does not come with the added complexity of choosing contraints for *post hoc* calculations.

Chapter 2

Integrating protein structure and sequence variation to identify sites under selective pressure

2.1 Introduction

This work has been previously published in the journal *Molecular Biology and Evolution*.[1] Many approaches to detect sites under selection aim to identify sites with evolutionary rate ratio $\omega = dN/dS$ significantly larger than one [70, 95]. Maximum-likelihood models are fit to sequence alignments and ω values for each site are estimated using either random effects (REL) or fixed effects (FEL) [51, 55, 70, 109]. Most REL models pre-specify a number of rate classes and fit the ω values for each class as well as the fraction of sites belonging to each class [51, 70, 109]. Rates for individual sites are recovered via an empirical Bayes approach. Some works have also attempted to determine the optimal number of rate classes, either via a goodness-of-fit criterion [55] or by employing a Dirichlet process which fits the number of rate classes as well as their properties [42, 85]. By contrast, FEL models directly fit an individual

[1] A. G. Meyer and C. O. Wilke. Integrating sequence variation and pro- tein structure to identify sites under selection. Molecular Biology and Evolution, 30:3644, 2013. C. O. Wilke helped to design the project and write the manuscript.

ω value to each site [51], thus allowing for as many different rates as there are sites in the sequence alignment.

One inherent limitation of all these approaches is that they cannot provide a baseline expectation for the ω value of a given site. For example, a site with $\omega = 0.9$ would not be identified as being under positive selection, yet $\omega = 0.9$ might be unusually high—and possibly indicative of selection for function—if the baseline expectation for this site in this protein was $\omega = 0.1$. Likewise, sites with particularly low ω—indicative of negative selection and likely functional importance—cannot be identified at all without a baseline expectation.

Here, we develop maximum-likelihood models that can provide a baseline expectation for ω and can identify sites that deviate from this baseline. Our method is based on the observation that the evolutionary conservation of a site is correlated with the site's relative solvent accessibility (RSA, a measure of solvent exposure of the focal amino acid in the folded, 3-dimensional protein structure) [9, 16, 29, 34, 68, 80, 86, 99]. In our models, ω is described by linear functions of RSA. We use a model-fit criterion to identify the optimal number of linear functions required to describe all sites in a protein, and for each site we identify to which linear function it most likely belongs.

For a test case, we apply our method to two viral proteins, influenza hemagglutinin and neuraminidase. We find that models in which ω is RSA-dependent always provide a better fit than conventional, RSA-independent models. Further, we find that the number of different linear functions needed to

describe these viral proteins is small, on the order of 6-10. In general, most sites in a protein fall into an RSA-dependent range of ω values that we consider the baseline expectation. Sites outside the baseline are candidates for functional selecton. In the case of hemagglutinin, these off-baseline sites are enriched in sites near the sialic-acid binding region. In the case of neuraminidase, few sites fall clearly outside the baseline region, with the exception of the well-known oseltamivir resistance site 274. Our method is easily implemented and broadly applicable to a wide range of scenarios, as long as a crystal structure is available for the protein of interest.

2.2 Materials and Methods

2.2.1 Sequence preparation

Sequences were downloaded for hemagglutinin 3 (H3) and neuraminidase (N1) from the influenza Research Database [90]. We selected human influenza A sequences including strains isolated from all geographic regions and years. The full set of H3 and N1 included over 10,000 nucleotide sequences from each protein. This set was then pared to remove all duplicated sequences. After processing, 2078 sequences remained for hemagglutinin and 3322 sequences remained for neuraminidase. Next, a protein structure was downloaded from the protein data bank (PDB) corresponding to each of the two proteins (PDBID: 1RD8 for hemagglutinin; PDBID: 1NN2 for neuraminidase). All nucleotide sequences were translated and aligned to the amino acid sequence from the corresponding PDB file, using the MUSCLE sequence alignment tool with de-

fault settings [28]. After alignment, gaps that were introduced relative to the sequence in the PDB file were removed, and the amino acids were reverted to their nucleotide codons. To make the subsequent evolutionary rate fitting more computationally tractable, we randomly selected 500 of the original pared sequences to be included in further analysis.

For both alignments of 500 sequences, we generated a phylogenetic tree with RAxML [91]. We used the GTRCAT substitution approximation available in RAxML; this approximation was chosen to make computing a phylogenetic tree computationally tractable with the large number of sequences used here. Similarly, the multithreading option was used in the RAxMLHPC version to speed up the computation.

2.2.2 RSA Determination and Binning

Hemagglutinin and neuraminidase are both functional multimers in solution. We used the crystal symmetry of their X-ray structures to determine the most likely multimeric form for each protein. The program DSSP version 1.0 [49] was used to calculate the solvent accessibility (SA) per site on both the monomeric and physiologically relevant multimeric forms, and the absolute accessibility was normalized as described previously [9, 98]. The sequence data was then subdivided into eight evenly-spaced bins according to the RSA of their sites in the protein structure, as described by Scherrer et al. [86].

2.2.3 Evolutionary rate determination

We implemented a variant of the Goldman-Yang codon evolution model (GY94, [35]) in the phylogenetic modeling language HyPhy [53]. The model we used is an extension of the model proposed by Scherrer et al. [86]. Briefly, we used the standard GY94 matrix but made the evolutionary-rate ratio $\omega = dN/dS$, the branch length t, and the transition to transversion ratio κ linear functions in relative solvent accessibility (RSA). We express their RSA dependence as:

$$\omega(\text{RSA}) = \omega_a \times \text{RSA} + \omega_b \,, \tag{2.1}$$

$$t(\text{RSA}) = t_a \times \text{RSA} + t_b \,, \tag{2.2}$$

$$\kappa(\text{RSA}) = \kappa_a \times \text{RSA} + \kappa_b \,. \tag{2.3}$$

Further, ω_a and ω_b are random effects, drawn from discrete distributions with a finite set of categories. Specifically, the distribution of ω_a is described by pairs of values (ω_a^k, p_a^k), such that $\Pr\{\omega_a = \omega_a^k\} = p_a^k$, where $k = 1, \ldots, D_a$, $\omega_a^k \geq 0$, $p_a^k \geq 0$, $\sum_k p_a^k = 1$. Similarly, the distribution of ω_b is described by pairs of values (ω_b^k, p_b^k), where $k = 1, \ldots, D_b$. The paramter D_a determines the number of ω-slopes in our model, and the parameter D_b determines the number of ω-intercepts. All other parameters (t_a, t_b, κ_a, κ_b) are fixed effects.

The infinitesimal matrix generator Q for the GY94 model has the usual

form (for $i \neq j$)

$$Q_{ij} = \begin{cases} 0 & \text{if more than one nucleotides changes} \\ \pi_j & \text{if synonymous transversion} \\ \kappa\pi_j & \text{if synonymous transition} \\ \omega\pi_j & \text{if nonsynonymous transversion} \\ \kappa\omega\pi_j & \text{if nonsynonymous transition} \end{cases}, \tag{2.4}$$

where π_j is the frequency of codon j, the indices i and j run over all 61 sense codons, and κ and ω are RSA-dependent, as stated above. The transition matrix for finite evolutionary time t becomes

$$P = e^{tQ}. \tag{2.5}$$

Here, the branch length t is also a linear function of RSA. Since t can be considered equivalent to the synonymous substitution rate, our model does not assume a single fixed synonymous substitution rate at every site, as is the case in the conventional GY94 model. Scherrer et al. [86] had previously found that models with RSA-dependent t and κ fit yeast data better than models with constant t and κ. We confirmed this observation here for influenza proteins.

2.2.4 Optimum model determination

We chose to implement a random-effects model [109] with independent slopes and intercepts. Optimization of fit parameters was performed by maximum likelihood estimation in HyPhy. All parameters except the codon frequencies π_j were determined by maximum likelihood. Estimated codon frequencies were calculated from the entire sequence alignment using the F3x4 model.

To identify the overall best-fitting model we used the Akaike Information Criterion (AIC) [2, 14]. We fit between zero and five slopes and between one and five intercepts in all pairwise combinations. We defined the best-fitting model as the one with the minimum AIC value.

2.2.5 Structural mapping

To assign sites to rate classes, we calculated posterior probabilities using the empirical Bayes approach [70]. For further analysis, we considered sites to either evolve at the rate given by the class with the highest posterior probability, or at an average rate calculated over all classes and weighted by the posterior probabilities.

For each of the proteins tested, sites were selected that showed a high ω and a low RSA or, conversely, a low ω and a high RSA. Sites of interest were mapped back to the original protein structure using the molecular visualization tool PYMOL [87]. For comparison, other predicted or verified sites were mined from the literature. In cases where the numbering convention used for the published sites was unclear, we three-dimensionally aligned the protein structure in the literature with our reference structure, using the PYMOL plugin CEALIGN. After three-dimensional alignment, corresponding sites could be mapped regardless of the numbering convention used for each protein.

2.3 Results

2.3.1 General approach

We introduced an RSA dependence to the evolutionary rate ω, the transition/transversion ratio κ, and the branch length t. RSA was incorporated by making ω, κ, and t linear functions of RSA. For example, ω with RSA dependence becomes $\omega = \omega_a \times \text{RSA} + \omega_b$, as described previously [86]. In the context of random-effect likelihood models, the intercept ω_b, the slope ω_a, or both can be either random or fixed effects. Note that Scherrer et al. [86] used fixed effects for both ω_a and ω_b. Here we systematically evaluated all possible combinations of fixed and random effects. We fit models with multiple intercepts and a single slope (ω_b is a random effect and ω_a is a fixed effect), with a single intercept and multiple slopes (ω_b is a fixed effect and ω_a is a random effect), and with both multiple intercepts and multiple slopes (both ω_b and ω_a are random effects). For comparison, we also fit a model with one intercept and one slope (both ω_b and ω_a are fixed effects). We treated κ and t as fixed effects in all models. We represented the traditional, RSA-independent approach by a random-effects model with multiple intercepts for ω and no slope for κ, t, or ω. In each case, we identified the optimal number of rate classes by AIC.

Conventionally, one would use an RSA-independent model to identify sites with $\omega > 1$. Including an RSA dependence allows us to identify a subset of sites with $\omega < 1$ that likely experience functional selection (e.g., sites that experience a combination of purifying and diversifying selection, such that the

resulting ω value falls below one but is still elevated relative to typical sites at this RSA). The RSA dependence also allows us to identify sites under particularly strong negative selection. Figure 2.1 shows a schematic representation of a typical result from our model. A canonical trapezoidal shape emerges from the data where we expect the vast majority of sites to be found. This shape is determined by the structure of the protein, and its exact location and size will vary among structures. Furthermore, if one performs a Box-Cox [11] transformation to produce normally distributed y-values, then dN/dS versus RSA becomes a rectangle with a simple RSA dependence. We consider any site outside of this region to be important and likely under selection (positive or negative) for function. As sites with $\omega > 1$ would be found by traditional analysis, additional sites found by our method are sites with low RSA and comparatively high $\omega < 1$ and sites with high RSA and very low $\omega \ll 1$.

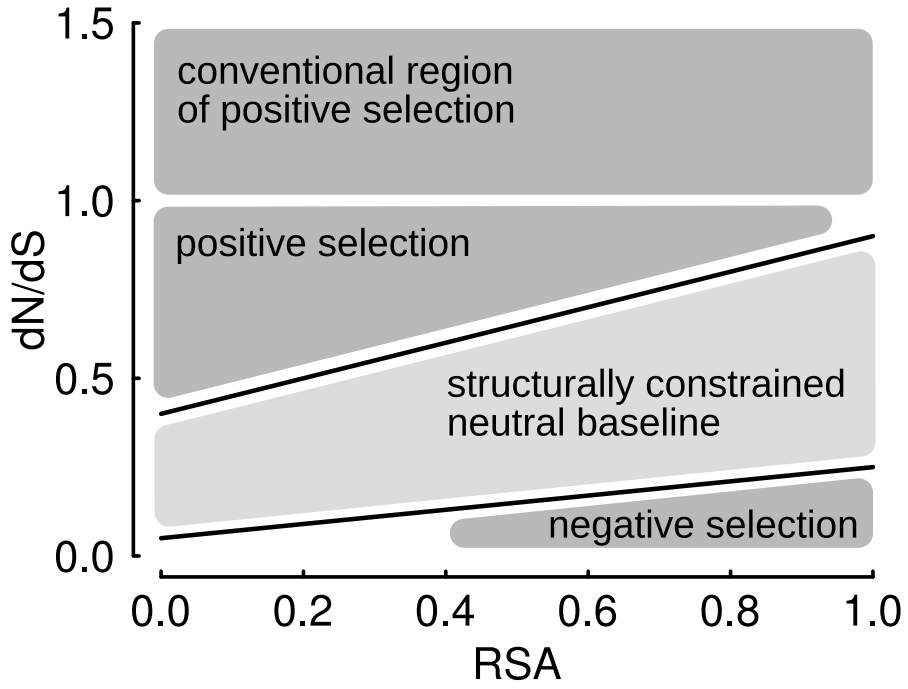

Figure 2.1: Regions of interest in ω–RSA plot. Most sites in proteins fall into a trapezoidal region we consider the neutral baseline. Sites with $\omega > 1$ are generally considered to be under positive diversifying selection. In addition to such sites, our method can also identify sites with an $\omega < 1$ but either larger or smaller than expected given their RSA. These sites fall into the triangular regions below $\omega = 1$ that are either above or below the neutral baseline. Sites in these regions experience either an accelerated or a reduced rate of evolution relative to the baseline and are likely to be functionally important.

Unexpectedly conserved sites are particularly difficult to find by traditional analysis. Sites with low ω are abundant, because most sites experience some negative selection pressure solely due to the requirement that the protein fold properly and be stable. With no baseline expectation for ω at each site, it

is unclear which of the sites with low ω are particularly important, for example because they are critical for function. As the majority of conserved sites have low RSA values, conserved sites with high RSA stand out as unusual. By incorporating RSA into models of protein evolution, we can identify sites under purifying selective pressure that would normally be missed. Similarly, with our method, sites that have a high RSA and correspondingly high ω (possibly exceeding 1 by a small amount) can be included in the null hypothesis (the neutral baseline in Fig. 2.1); such sites might otherwise be considered to be undergoing positive selection.

2.3.2 Application to influenza hemagglutinin and neuraminidase

We applied our method to the influenza proteins hemagglutinin 3 (H3) and neuraminidase 1 (N1). We fit a total of 30 different models to each protein; in these models, the number of intercept classes varied between one and five, and the number of slope classes varied between zero and five. The models were compared directly by subtracting the AIC of each model from the global minimum AIC (corresponding to the best-fitting model). For hemagglutinin, the model containing three slopes and three intercepts provided the global minimum AIC (Figure 2.2). For neuraminidase, the global minimum was found in the model with two slopes and three intercepts (Figure 2.3). Furthermore, all models with at least one slope (i.e., incorporating RSA) gave a substantially better fit than those not incorporating RSA. Likelihood values and numbers of fitted parameters for all models are given in Tables S1 (for hemagglutinin)

and S2 (for neuraminidase).

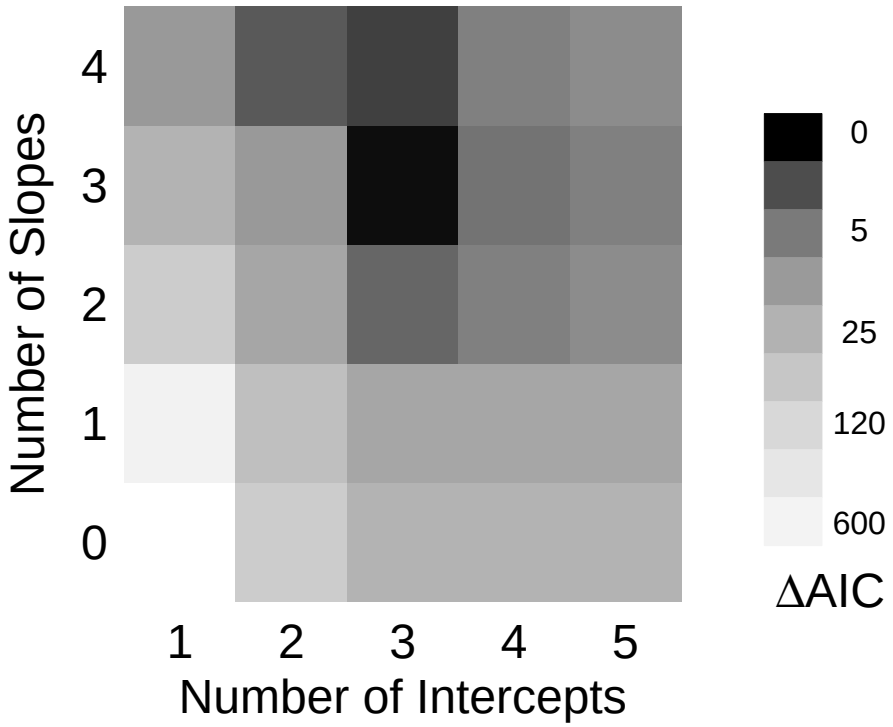

Figure 2.2: Model fit as a function of the number of slopes and intercepts in the model, for the influenza haemagglutinin trimer. The shading reflects the difference in AIC between the best model (3 slopes and 3 intercepts in this case) and all other models.

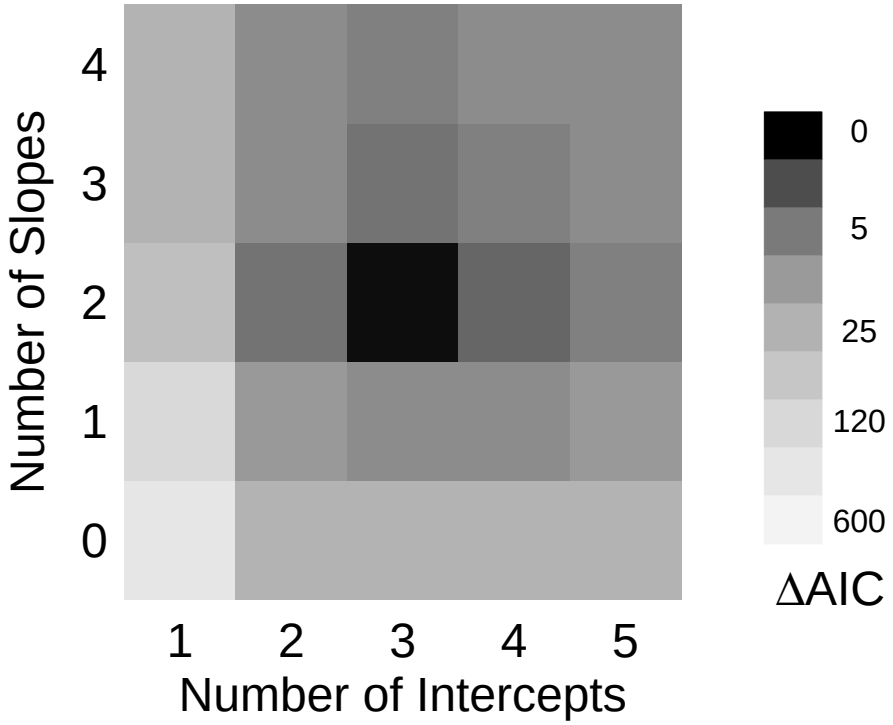

Figure 2.3: Model fit as a function of the number of slopes and intercepts in the model, for the influenza neuraminidase tetramer. The shading reflects the difference in AIC between the best model (2 slopes and 3 intercepts in this case) and all other models.

Hemagglutinin forms a homo-trimer on the surface of quiescent influenza virus, and neuraminidase is only enzymatically active as a homo-multimer (tetramer) within infected cells. Therefore, for both proteins it is not clear whether RSA measurements should be performed on the monomeric or the multimeric forms. It seems likely that both the monomeric and multimeric forms will have some influence on sequence evolution. For example, the monomeric form of a protein should evolve to prevent unfavorable homo-

multimerization. By contrast, the multimer must evolve to form a stable quaternary structure. Both the monomeric and multimeric forms will evolve to prevent aggregation, but they may do so on different surfaces. Further complicating matters, even for proteins that function only as multimers we generally have no information about the fraction of time they spend as multimers. If, in solution, the proteins spend very little time multimerized, evolution due to biophysical constraints will act more often on the monomeric forms. We calculated RSA values on both the monomeric and the multimeric forms for both hemagglutinin and neuraminidase. For both proteins, the multimeric form provided a better fit than the monomeric form for every combination of slopes and intercepts (not shown). Therefore, we considered only the multimeric forms of these proteins for further analysis.

After fitting the models, we assigned individual evolutionary rates to each site. We employed the empirical Bayes approach to calculate posterior probabilities for each site to belong to each rate class. We then considered two alternatives of how to convert rate classes into site-specific rates. First, we assigned sites to their most probable slope and intercept and then calculated evolutionary rates given each site's RSA. Second, we calculated an average rate by weighting each rate class with the probability that a site falls into that class. To calculate the weighted average ω_{ave}, we write

$$\omega_{\text{ave}} \quad = \quad \sum_{i,j} pp_{ij}(\omega_a^i \times \text{RSA} + \omega_b^j)\,, \tag{2.6}$$

where ω_a^i is the slope in category i, ω_b^j is the intercept in category j, and pp_{ij}

is the posterior probability of slope i and intercept j at the site of interest.

Figure 2.4 shows results for hemagglutin obtained under the first approach. We selected four representative cases. The top left graph represents the traditional, RSA-independent case. The top right and bottom left graphs represent the mixed models with a single slope and multiple intercepts and with multiple slopes and a single intercept, respectively. The bottom right represents the overall best-fitting model, with $\Delta AIC = 0$. Figure 2.5 shows the same four cases with rates calculated under the averaging scheme. In comparison to Figure 2.4, averaging reduces some of the very high ω values. Note also that ω_{ave} values are less sensitive to the exact model specification. For the four models shown in Figure 2.5, the ω_{ave} values for the best model correlate with the ω_{ave} values for the other three models with Spearman $\rho = 0.91$, $\rho = 0.94$, and $\rho = 0.91$, respectively (in order top left, top right, bottom left). All correlations are highly significant.

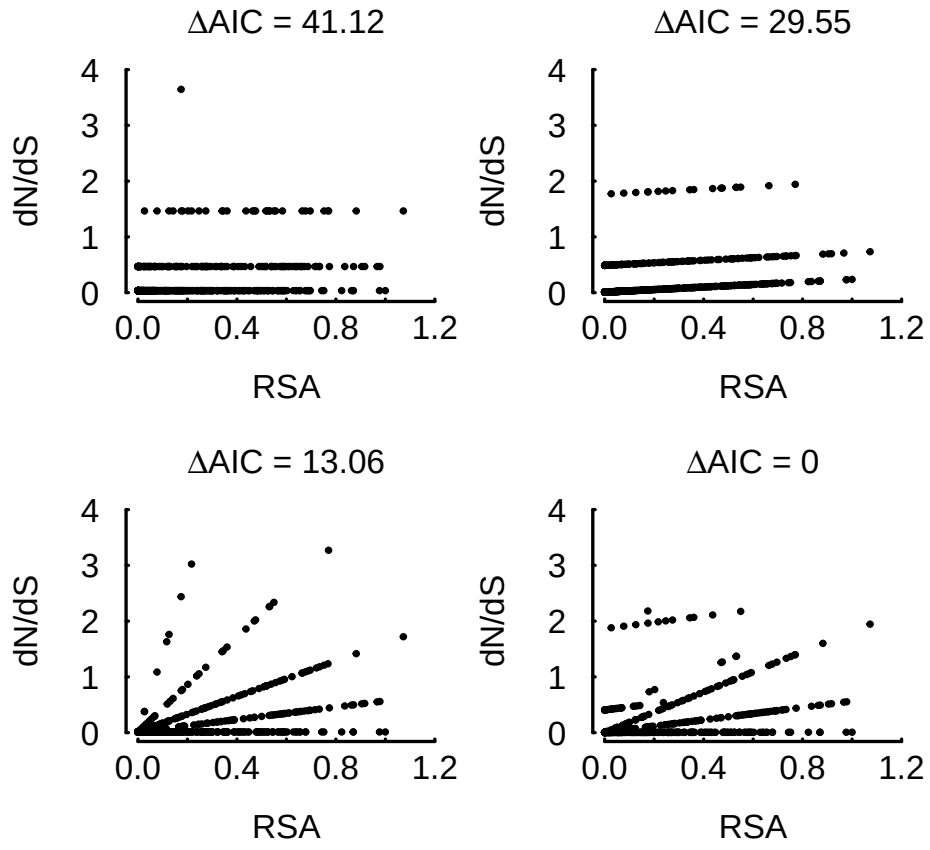

Figure 2.4: Assignments of sites to rate classes, for the influenza haemagglutinin trimer. Each graph shows each site's dN/dS plotted against the site's RSA. Sites are assumed to evolve at a dN/dS determined by the rate class they are most likely to fall into. Top left: The best model with multiple intercepts and no slope (no RSA dependence). Top right: The best model with multiple intercepts and one slope. Bottom left: The best model with multiple slopes and a single intercept. Bottom right: The overall best model, with three intercepts and three slopes. ΔAIC values are calculated relative to the overall best model. Figure 2.5 shows the same results but averaged over rate classes.

By fitting an RSA-dependent model to hemagglutinin data, we could identify sites with high ω that may not experience accelerated evolution. The

RSA-independent model (Figure 2.5, top left) showed that there are a number of sites with $\omega > 1$. By incorporating RSA, we were able to filter out those sites with high ω that had correspondingly high RSA. For hemagglutinin, the best-fitting model suggested that at least one site with elevated evolutionary rate should be considered part of the neutral baseline (Figure 2.6). We also found a set of very exposed sites that were highly conserved. In total, for hemagglutinin we found 33 sites that we predicted to experience accelerated evolution (above the upper dashed line in Figure 2.6) and 9 sites that we predicted to be exceptionally conserved (below the lower dashed line in Figure 2.6). These sites are listed in Table S3.

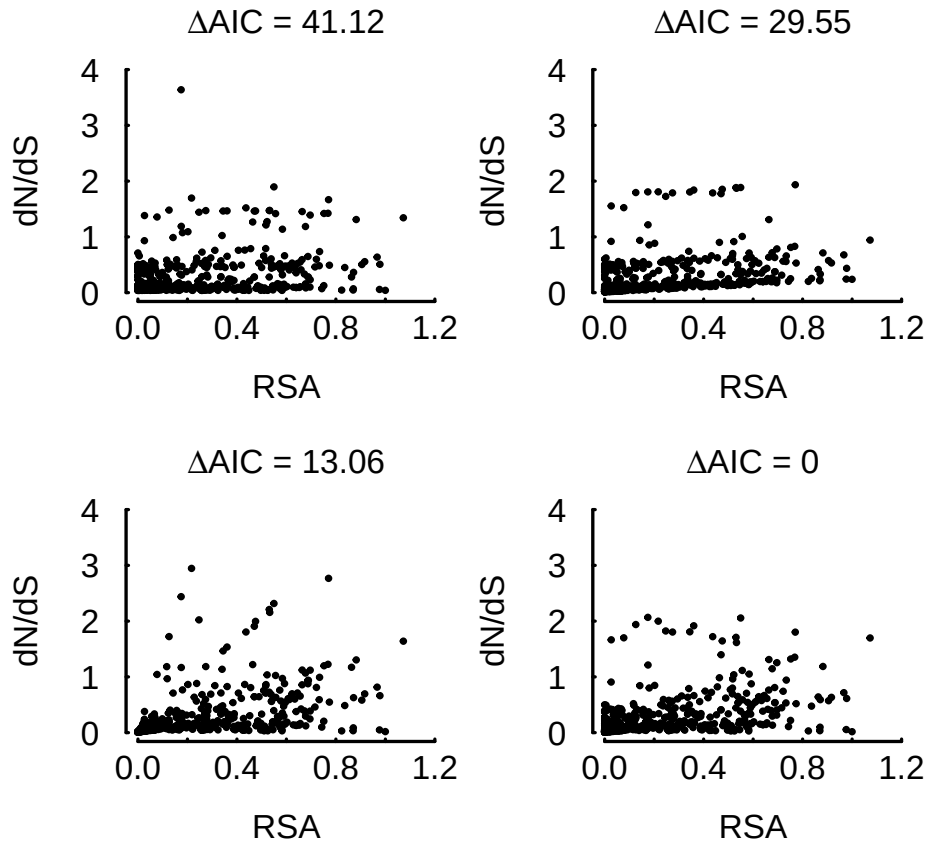

Figure 2.5: Assignments of sites to evolutionary rates, for the influenza haemagglutinin trimer. Each graph shows each site's weighted average dN/dS plotted against the site's RSA. Top left: The best model with multiple intercepts and no slope (no RSA dependence). Top right: The best model with multiple intercepts and one slope. Bottom left: The best model with multiple slopes and a single intercept. Bottom right: The overall best model, with three intercepts and three slopes. ΔAIC values are calculated relative to the overall best model.

Hemagglutinin is an important target for protective immunity and has therefore been the topic of many previous analyses. It is well known that many sites in human-influenza hemagglutinin experience adaptive evolution,

in particular in domain 1 of the protein, which contains the sialic-acid binding site [8]. Here, we asked whether sites in the sialic-acid binding region were enriched in sites experiencing accelerated evolution. The sialic-acid binding site is expected to be under positive selection, as it is a major target for host range shifts and antibody binding [103]. We identified all sites within 8Å of sialic acid in the hemagglutinin structure, and analyzed where they fell in the ω_{ave}–RSA plot. We found 39 such sites. (Note that we do not expect all of these sites to have high ω, and neither do we expect all sites with high ω to be near the sialic acid binding region.) Of these 39 sites, 10 fell above the dashed line in Figure 2.6 and 29 below. This represents a significant positive enrichment of such sites at high ω (Fisher's exact test, OR $= 6.6$, $p = 6.1 \times 10^{-5}$).

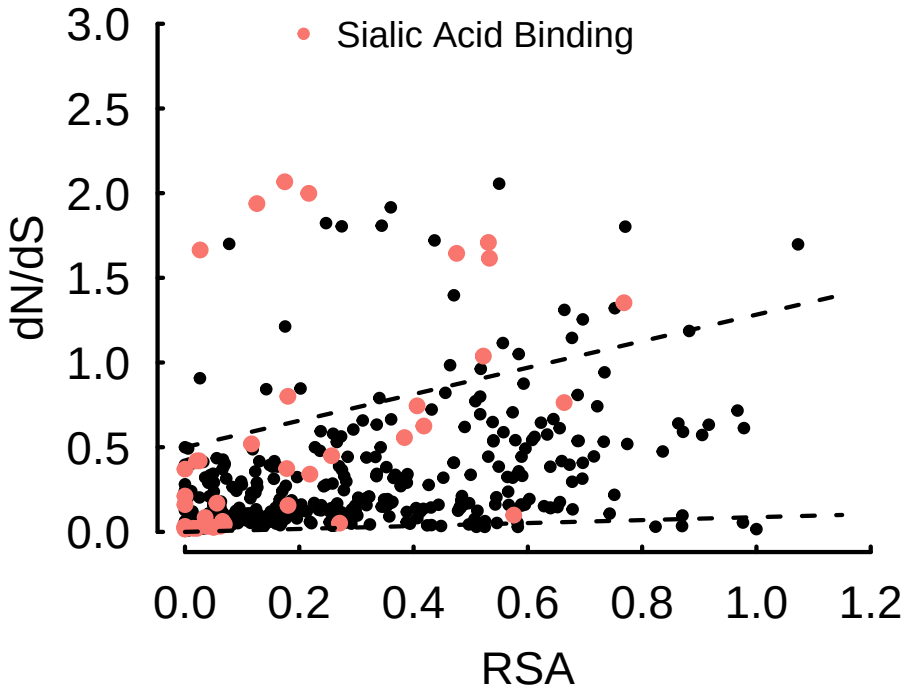

Figure 2.6: Average ω versus RSA for hemagglutinin, obtained from the optimal model (3 slopes and 3 intercepts). Dashed lines indicate the trapezoidally shaped neutral baseline (as ascertained by eye). Sites highlighted in red are within 8Å of the sialic-acid binding region. Sites above the upper dashed line are significantly enriched in sites near the sialic-acid binding region (Fisher's exact test, OR $= 6.6$, $p = 6.1 \times 10^{-5}$).

We visualized all sites evolving at either accelerated or reduced ω by mapping them onto the hemagglutinin structure (Figure 2.7). We found that the majority of the sites with accelerated ω fell near the top of the hemagglutinin structure, but others occurred in small clusters throughout the structure (Figure 2.7A). The clustering of these remaining sites suggests that their evo-

lutionary rate is driven by selection pressure mediated by antibody binding. The sites with unusually low ω given their RSA fell into three categories. Three sites (69, 121, 141) were relatively close to the sialic-acid binding region (Figure 2.7B, C), even though not within the 8Å we used as cutoff to identify sites near the sialic-acid binding region. (Their distances to the sialic-acid binding region were 17Å, 18Å, and 12Å, respectively). These sites may provide crucial structural support for proper functioning of the hemagglutinin protein. Three more (163, 236, 238) seemed to be involved in the trimer interface (Figure 2.7B, C). The remaining sites were located throughout the protein, and their possible function was not readily apparent (Figure 2.7A).

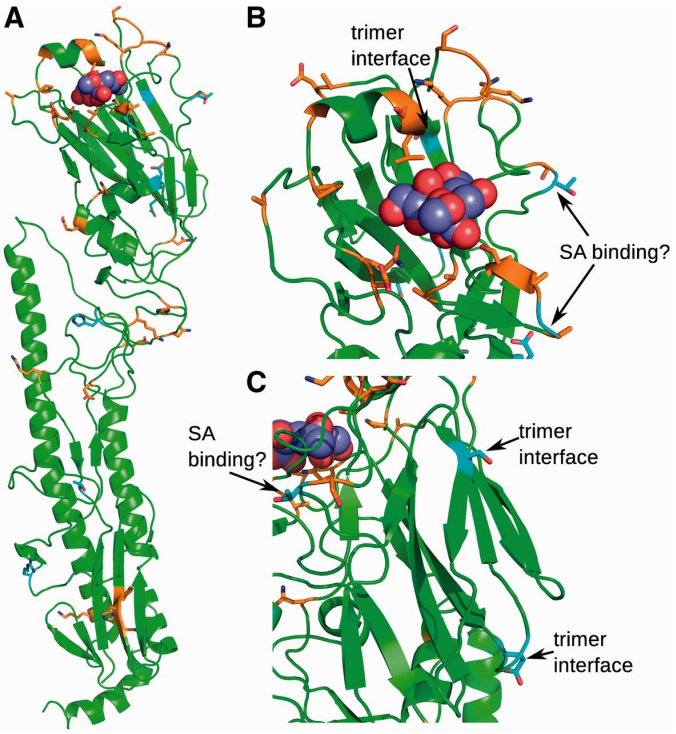

Figure 2.7: Sites of interest identified for hemagglutinin. Sites that fall above the upper dashed line in Fig. 2.6 are colored orange. Sites that fall below the lower dashed line in Fig. 2.6 are colored light blue. The polypeptide backbone is colored green. Sialic acid is represented by the space filling model near the top of the molecule. (A) View of the entire hemagglutinin monomer. (B) View of the sialic-acid binding region. Sites that are highlighted as "SA binding?" are unusually conserved and close to (though not within 8Å of) the sialic acid. Sites that are highlighted as "trimer interface" are unusually conserved and seem to be important for trimerization. (C) View of the trimer-interface region. Labeling of sites is as in part (B).

We next considered neuraminidase. Overall, neuraminidase showed substantially lower ω values than hemagglutinin. While hemagglutinin had many sites with $\omega > 1$, neuraminidase had just a few. The model with multiple intercepts and no slopes placed five sites at $\omega > 1$ (Figures 2.8 and 2.9,

top left). No individual site seemed to stand out particularly under the RSA-independent approach. By contrast, under our best model for neuraminidase, with 2 slopes and 3 intercepts, one site stood out as having low RSA and particularly high ω (Figures 2.8 and 2.9, bottom right). This site is position 274; a common oseltamivir resistance mutation occurs at this site (most commonly H274Y). For neuraminidase, we did not find any highly exposed sites that were particularly conserved. In total, for neuraminidase we found 9 sites that we predicted to experience accelerated evolution (Table S3).

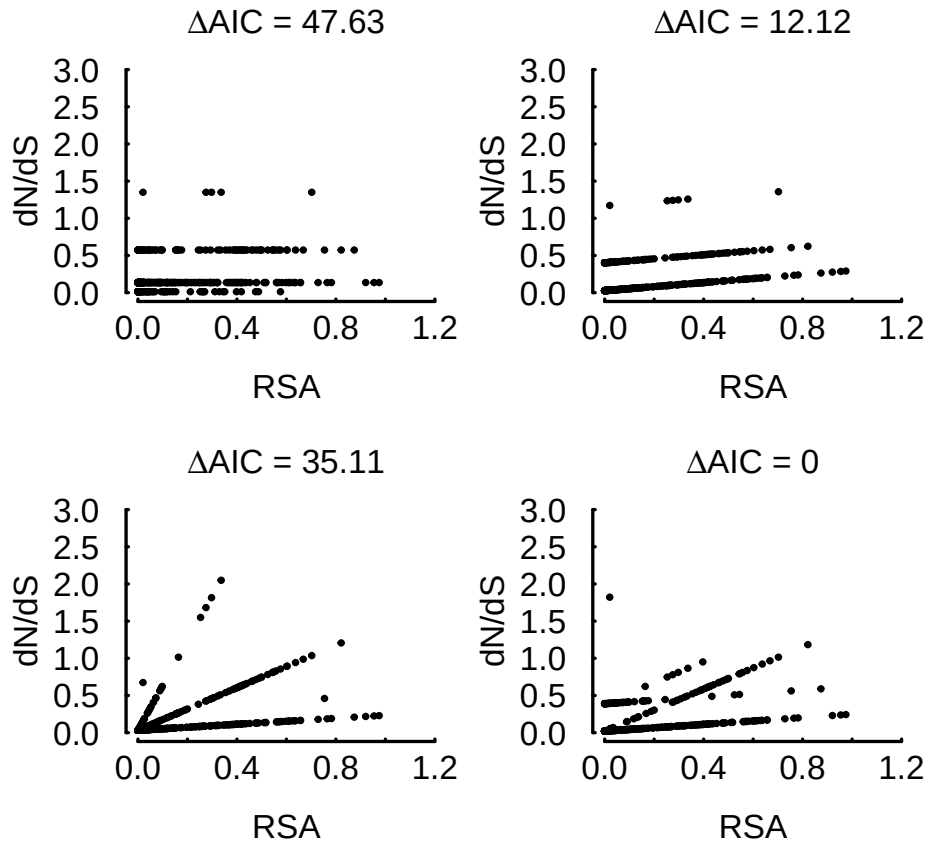

Figure 2.8: Assignments of sites to rate classes, for the influenza neuraminidase tetramer. Each graph shows each site's dN/dS plotted against the site's RSA. Sites are assumed to evolve at a dN/dS determined by the rate class they are most likely to fall into. Top left: The best model with multiple intercepts and no slope (no RSA dependence). Top right: The best model with multiple intercepts and one slope. Bottom left: The best model with multiple slopes and a single intercept. Bottom right: The overall best model, with three intercepts and two slopes. ΔAIC values are calculated relative to the overall best model.

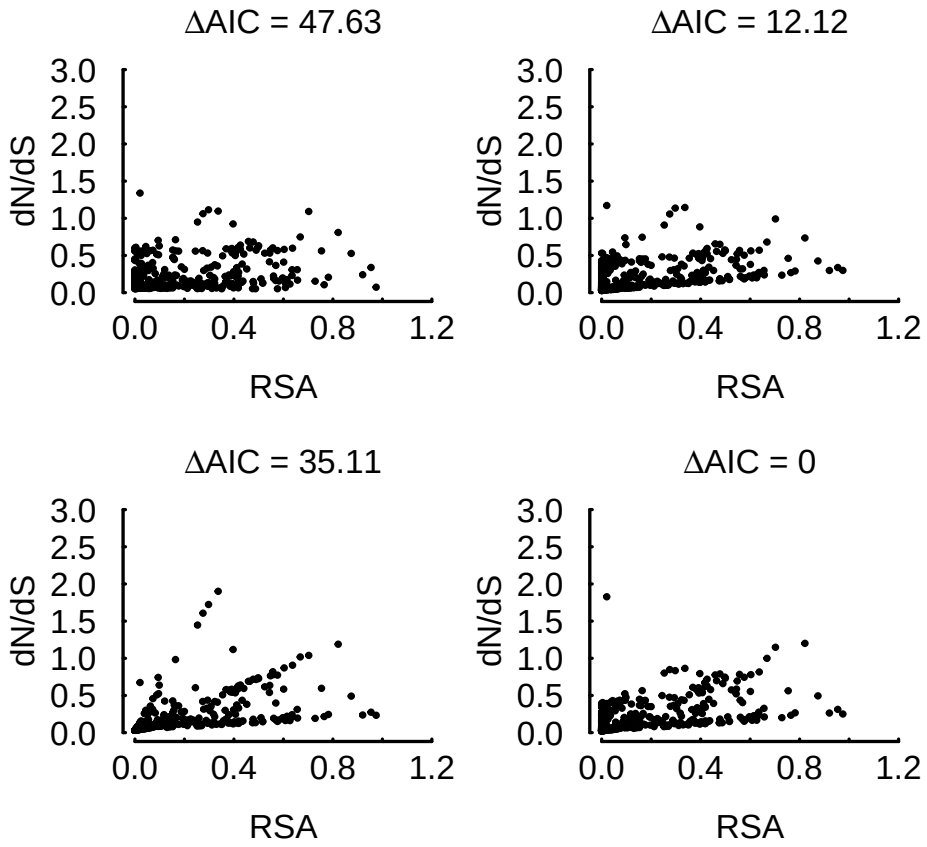

Figure 2.9: Assignments of sites to rate classes, for the influenza neuraminidase tetramer. Each graph shows each site's weighted average dN/dS plotted against the site's RSA. Top left: The best model with multiple intercepts and no slope (no RSA dependence). Top right: The best model with multiple intercepts and one slope. Bottom left: The best model with multiple slopes and a single intercept. Bottom right: The overall best model, with three intercepts and two slopes. ΔAIC values are calculated relative to the overall best model.

2.3.3 Comparison of identified sites with prior work

The positively sites we found here were broadly in agreement with previously identified sites. In a seminal paper, [15] used positively selected sites in H3 to predict influenza evolution. Of the 18 sites they found, 11 sites fell above our null expectation (Figure 2.10, left). More recently, [56] identified 24 sites under directional selection in hemagglutinin; we identified 11 of those 24 sites (Figure 2.10, right). Both studies also identified a few sites that had a very low ω in our analysis. Some of the differences between our and their results are likely due to differences in the sequences analyzed.

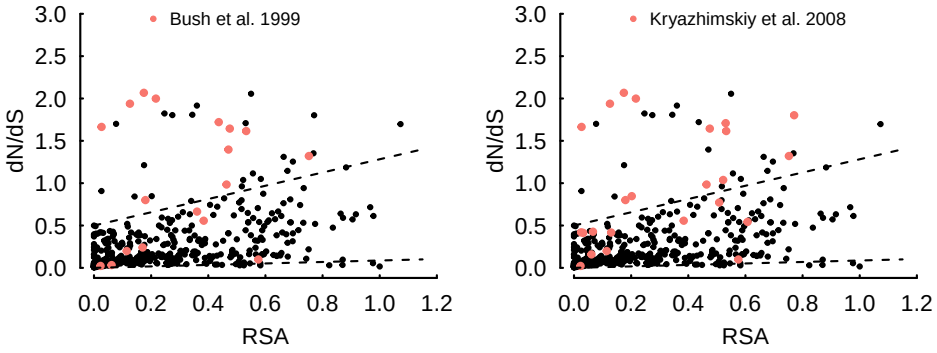

Figure 2.10: Comparison of our results with previous work on hemagglutinin. The left graph highlights sites found by [15] to be under positive selection. The right graph highlights sites found by [56] to be under directional selection. Dashed lines indicate the putative region of the structurally constrained neutral baseline.

Neuraminidase is the protein responsible for enzymatic cleavage of sialic acid following viral entry. In comparison to hemagglutinin, neuraminidase is a less commonly studied protein due, in part, to its intracellular function. As

neuraminidase is never exposed to the periplasm, it is not likely to be a major target of protective immune responses. Therefore, we do not expect many of the sites in neuraminidase to evolve very rapidly. Indeed, most of its sites show very high conservation; as mentioned previously, neuraminidase contains very few sites with $\omega > 1$ (Figures 2.8 and 2.9).

Of those sites found previously by [10], we found only the oseltamivir resistance site (site 274 in the PDB structure 1NN2, site 275 in [57]) to be under positive selection (Figure 2.11, left). Those Bloom sites conferring epistatic stability to the resistance mutation were generally elevated relative to a regression line, but they did not fall significantly above of the baseline expectation.

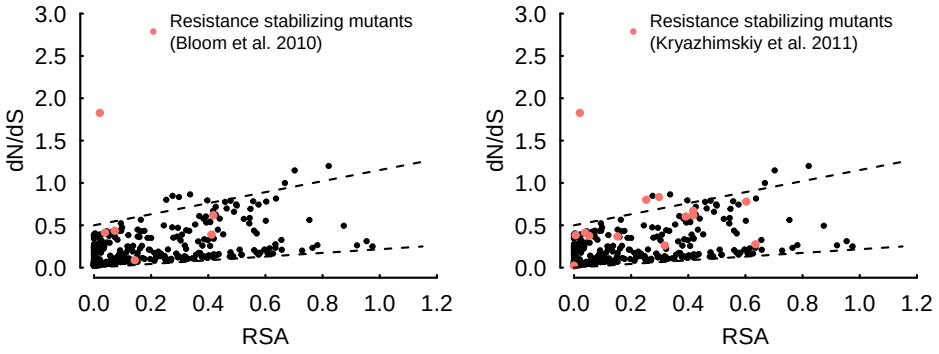

Figure 2.11: Comparison of our results with previous work on neuraminidase. Left: Sites found by [10] to be involved in the evolution of oseltamivir resistance are highlighted in red. Right: Site 274 and sites found by [57] to have 274 as trailing site are highlighted in red.

We also compared our results to recent work by [57] (Figure 2.11, right). They identified sites in neuraminidase that enabled subsequent substitutions at other sites. We specifically considered sites that they identified as leading to

substitutions at the oseltamivir resistance site. Among these sites were half of the sites identified by [10], so these two studies give highly congruent results. We found two sites from [57] significantly above the baseline expectation. Our sites congruent with [57] are 220 and 430 (in PDB structure 2HU0), both near the sialic-acid-cleaving active site. The remaining sites from [57] were also generally elevated, but did not rise above the baseline expectation.

2.4 Discussion

We have described a new method for identifying important sites in protein-coding sequences. Furthermore, we have shown that this new method fits molecular sequence data better than a conventional random-effects likelihood (REL) approach. To generate improved models, we used the correlation between RSA and evolutionary rate [29] to define a new, more accurate evolutionary null expectation. We applied our method to two influenza proteins, hemagglutinin 3 and neuraminidase 1; we classified sites with ω greater than expected given their RSA as positively selected, and sites with ω lower than expected as negatively selected. Sites found by our method are in agreement with experimentally validated sites and with results from other computational studies. Our method goes beyond previous approaches by finding sites that would previously have been missed (reducing Type II errors) and by including some sites in the null that would normally have been rejected (reducing Type I errors).

Several approaches to fitting evolutionary rate have been developed

and improved in the last two decades [35, 51, 52, 55, 69, 70, 85, 95, 107, 109]. One approach, fixed-effect-likelihood (FEL) based fitting [51], involves estimating an independent ω value at each site; although a FEL approach limits Type I errors, it can lead to model over-parameterization as there is no obvious way to penalize over-parameterized models. Another approach, counting methods [51], is conceptually simpler but estimates independent κ and t parameters at each site. While ω may vary substantially among sites, κ and t are not likely to do so; therefore, estimating them per site leads to overfitting. Moreover, as sites are treated independently, both FEL and counting methods require a large number of sequences per gene to gain power [51]. Traditional REL-based rate estimation relies on a predefined distribution for ω and a pre-specified number of rate classes [51]. As a result, all sites are aggregated to increase power when a large number of sequences are not available; however, with a very small number of sequences, REL methods tend to lead to increased Type I errors [51]. In addition, estimating the number of rate classes becomes a potential complicating factor for REL methods: Even though a model with k classes is nested into a model with $k+1$ classes, one cannot simple carry out a likelihood ratio test to determine which model provides the better fit, because the nesting is not identifiable [3].

Here, we chose to identify the optimal number of classes by testing all plausible candidate REL models and ranking them according to their AIC. We found that the total number of classes needed was relatively small: two to three slopes and three intercepts were sufficient to describe relatively large

alignments of hemagglutinin and neuraminidase. In general, the number of classes for optimal fit will likely vary amongst proteins. Alternative methods of identifying the optimal number of rate classes include adding classes until a goodness-of-fit criterion fails [55] or using a Dirichlet process to fit number of classes as a parameter in the model [42, 85]. Either of these approaches could be employed with our RSA-dependent model of codon evolution.

When using a random-effects model, one has to decide how to assign ω values to individual sites. We considered two alternative approaches. To each site we assigned either the ω value corresponding to the most probable rate class for the site or an average ω value calculated as a weighted average over all classes. We think that both approaches are valuable. The first approach produces simpler graphs, as it highlights the linear RSA dependency employed in the model. Using this approach, we can easily tell which sites fall into common rate classes (i.e., are part of the neutral baseline) and which sites have unusual ω given their RSA. A downside of this approach is that sites which have comparable posterior probabilities for two or more classes will appear to belong exclusively to a single class. This downside is alleviated by the second approach, which will place those sites at intermediate, averaged ω values. A second advantage of the averaging approach is that average ω values are relatively insensitive to the model specification (exact number of slopes and intercepts).

We analyzed a Goldman-Yang model [35] in which both the evolutionary-rate ratio ω and time t were RSA-dependent, but only ω was implemented

as a random effect that could vary among sites independently of RSA. An alternative model would be a structure-aware Muse-Gaut model [69]. The Muse-Gaut model uses parameters α and β instead of ω and t, with $\alpha = t$ and $\beta = \omega t$. In models that incorporate site variability but no structural information, the Muse-Gaut model is generally preferred over the Goldman-Yang model [25, 52, 55], because the Muse-Gaut model naturally allows for variation in both synonymous and non-synonymous evolutionary rates (by treating both α and β as random effects). By contrast, when RSA is taken into account but other site variation is ignored, the Goldman-Yang model performs better than the Muse-Gaut model, indicating that ω varies linearly with RSA and β does not [86]. Whether the Muse-Gaut or the Goldman-Yang formulation should be preferred in the most general case, incorporating both structural information and synonymous and non-synonymous rate variation, remains an open question for future research.

We have shown that accounting for biophysical constraints in models of sequence evolution can reduce the frequency of Type I errors, by excluding sites with $\omega \sim 1$ at high RSA. For hemagglutinin, we were able to exclude at least one site with high evolutionary rate that had proportionately high RSA values. Our model likely also reduces Type II errors, by identifying sites with $\omega < 1$ as experiencing accelerated evolution. Several of the sites identified by [15] or [56] (Figure 2.10), or of the sites near the sialic-acid binding region (Figure 2.6), fell into this category. One drawback of our method is that defining which sites fall into the null expectation and which fall outside it

lies ultimately with the human researcher carrying out the analysis. Since our method is fundamentally prospective, geared towards identifying sites of potential interest in further experimental analysis, we do not consider this drawback as particularly severe. It is also important to keep in mind that sites with ω above the baseline but below 1 could either be subject to a mixture of positive and negative selection pressures or they could simply be subject to very little selection pressure at all, despite of their location in the protein. Finally, we saw some discrepancies between the sites we identified and sites found in previous works (Figure 2.10). We have no reason to believe that the inclusion of RSA into our model caused these discrepancies, since they also arose in our RSA-independent model. Most likely, these discrepancies were caused by differences in the alignments analyzed.

Several previous papers have proposed models of coding-sequence evolution that incorporate structural information. The simplest approach is to partition coding sequences according to a structural property (e.g., partition all sites into buried and exposed sites) and fit a fixed-effects model whose model parameters may vary by partition [6, 21, 86, 108]. These partitioned models tend to fit sequence data better than non-partitioned models. They are good at identifying differences among partitions, but they cannot identify individual sites that show unusual rates given their partition. More sophisticated models contain selection terms incorporating solvent accessibility or energetic interactions among residues in a structure [20, 81–84]. These models also tend to fit sequence data better than comparable models without struc-

tural information. However, it is not clear whether they can be used in a straightforward manner to identify individual sites under selection. The advantage of our method is that it is straightforward to implement, it requires only moderate amounts of processing time, and it produces results that are easily interpretable. A limitation of all these methods, including ours, with respect to viral evolution is that they assume site independence, when sites in most viral proteins are actually tightly linked. This limitation may cause overinflated variability in evolutionary rates or even biased estimates. The exact consequences of this limitation have received little attention and remain poorly understood, however.

One complicating factor in our approach comes from the proper measurement and interpretation of physiologically relevant protein structures. We tested both the monomeric and multimeric form of each protein, and found that the multimeric forms produced a better model fit in both cases. It is unclear whether the multimeric RSA would always be preferred. Goodness-of-fit may be related to the most common state of free, solvated protein. RSA values may also be incorrect because crystal structures reflect a single protein conformation but solvated proteins fluctuate among neighboring conformations. In future work, one could attempt to improve the accuracy of RSA calculations by estimating these fluctuations using molecular dynamics simulations.

Even though we introduced RSA dependence into the rate parameters of our model (ω, t, κ), we held the equilibrium codon frequencies π_i constant throughout all sites in the protein. Holding codon frequencies constant is a

convenient approximation that is usually employed when calculating evolutionary rates. However, in the structural context, it would be desirable to make codon frequencies depend on structure as well. Amino-acid frequencies vary with RSA, the most hydrophobic amino acids being the most common at low RSA values and the most hydrophilic ones being the most common at high RSA values [76, 80]. Consequently, codon frequencies must similarly vary with RSA. We suspect that some of this variation was absorbed into κ and t in our model. Nevertheless, a more realistic model would be desirable. At present, however, we are not sure how to formulate such a model. We could make the π_i linear functions of RSA as well, but this modeling choice would add a large number of parameters and possibly lead to an over-parametrized model.

In summary, our work has shown that accounting for RSA in evolutionary-rate models improves model fit. In addition, we were able to find the best-fitting model by exhaustively testing different numbers of slopes and intecepts. We could confirm extensive selection pressure near the sialic-acid binding site in the influenza protein hemagglutinin. Further, we could show that the oseltamivir-resistance site 274 in influenza neuraminidase stands out among the sites in this protein as experiencing particularly strong positive selection. Finally, our analysis of hemagglutin and neuraminidase offers new, potentially important sites for experimental investigation.

Chapter 3

Cross species correlation of evolutionary rate variation

3.1 Introduction

This work was previously published in the journal *Philosophical Transactions of the Royal Society B*.[1] Viral proteins are highly variable at the sequence level; they accumulate amino-acid substitutions at a rapid pace [27,77]. Yet their structures tend to be fairly conserved. Highly variable surface regions notwithstanding, most viral proteins need to maintain a specific structure to carry out their function in the viral life cycle (see e.g. [110]). The generally accepted picture is that sites in the protein core maintain the overall protein structure and are therefore most conserved. Sites on the surface are less critical to the protein structure and hence more free to vary, for example in response to selection pressures imposed by immune response. This view is based on the finding, replicated in widely differing organisms and using many different techniques, that on average sequence variability increases the closer a site is located towards the surface of a protein [9,16,20,24,29,34,36,68,86,97]. More

[1] A. G. Meyer, E. T. Dawson, and C. O. Wilke. Cross-species comparison of site-specific evolutionary-rate variation in influenza hemagglutinin. Philosophical Transactions of the Royal Society B, 368:1614, 2013. C. O. Wilke helped to design the project and write the manuscript.

specifically, in influenza, exposed sites in hemagglutinin and neuraminidase have been found to evolve faster than buried sites in these proteins [8, 66].

Thus, prior work has clearly established that protein structure influences site variability. What is less clear, however, is the magnitude of this effect. Is knowing a site is buried sufficient to predict that the site will be evolutionarily conserved, or are other factors stronger driving forces for site-specific evolutionary rates? And similarly, will homologous sites in related but distinct viral strains evolve at similar rates, or do the nature of the viral strain and the infected host organism impose stronger influences on site-specific evolutionary rates than the location of a site in the protein structure?

In this chapter, we address these questions for influenza hemagglutinin (HA). We compare per-site sequence evolution for two different host species (human and avian) and three HA subtypes (H1, H3, H5), and ask the following questions: (1) To what extent is rate variation determined by the location of a site in the structure, as measured by the sites' relative solvent accessibility? (2) To what extent is rate variation conserved within HA subtypes among viruses infecting different host species? (3) Are $\omega = dN/dS$ ratios elevated near the active site (the sialic-acid binding region) of HA? We find that protein structure, HA subtype, and host biology all affect rate variation in influenza HA.

3.2 Materials and Methods

3.2.1 Sequence preparation

We obtained sequences for hemagglutinin (HA) subtypes H1, H3, and H5 for human and avian hosts from the Influenza Research Database [90]. Using the built-in curating tools of the database, we carefully selected subsets of sequences that corresponded as much as possible to well-defined and distinct viral populations. Sequences were curated within each host species depending on its subtype. In particular, for each combination of HA subtype and host species, we considered only sequences that could be linked to a specific neuraminidase subtype.

Human H1 sequences were obtained from H1N1 strains isolated between 1977 (after the Fort Dix outbreak) until 2008 (before the 2009 flu pandemic). H1N1 strains since 2009 are not direct descendants of H1N1 strains before 2009 and thus were excluded. We found 2057 distinct H1 sequences. Human H3 sequences were obtained from H3N2 strains isolated between 1968 until 2012. We found 8315 distinct sequences. Human H5 sequences were obtained from H5N1 sequences without date restriction. We found 297 distinct sequences.

Avian sequences were curated by subtype with no restrictions placed on the date range; full data sets from FluDB of H1N1, H3N2, and H5N1 sequences were used. We found 106, 115, and 2684 distinct sequences, respectively.

In general, the process outlined in subsection 2.2.1 was used here. To align sequences and map them to the hemagglutinin structure, we downloaded

a reference structure from the protein data bank (PDB). We used the structure with PDB identifier 1rd8, which is a structure of the H1 subtype [92]. All nucleotide sequences were translated into amino-acid sequences. Each distinct set of sequences (human H1, human H3, and so on) was then aligned to the reference amino-acid sequence from the PDB file, using the MUSCLE sequence-alignment tool with default settings [28]. For alignments that contained more than 200 sequences, we selected a random subset of 200 sequences for further analysis. This reduction in alignment sizes was necessary because maximum-likelihood fitting of codon-evolution models becomes prohibitively slow for much larger alignments. In addition, phylogenetic trees were built just as in subection 2.2.1.

To assess the phylogenetic relationship between human and avian alignments, we also constructed combined human/avian alignments for H1, H3, H5 and reconstructed trees for the combined alignments. For H1, we found nearly complete phylogenetic separation between human and avian sequences. All but three avian and all but two human sequences formed a separate clade each. The remaining five sequences (introduced from classic swine lineage) were closer to each other than to either of the two clades. For H3, human and avian sequences were mostly separated. However, 14 avian sequences fell into the human clade (12 of those avian sequences grouped together as a single clade, introduced from the triple-reassortant H3N2 swine lineage). For H5, we found no phylogenetic separation. Human and avian sequences grouped together throughout the entire phylogeny. Human and avian H5 viruses are

not distinct lineages because human H5 is not transmitted effectively from human to human. All phylogenetic trees are provided as part of the online supplementary materials.

We calculated relative solvent accessibility (RSA) as described [66] and in model development. In brief, solvent accessibilities (SAs) were calculated with the program DSSP [49] and then normalized [9]. The sequence data was then subdivided into eight evenly-spaced bins according to the RSA of their sites in the protein structure, as described [86].

3.2.2 Evolutionary rate determination

We calculated site-specific evolutionary rates using two approaches that integrate sequence data and protein structure, a random-effects likelihood (REL) and a fixed-effects likelihood (FEL) approach. Both approaches were implemented in the phylogenetic modeling language HyPhy [53], and our HyPhy scripts are provided as part of the supplementary materials.

The REL approach was previously described [66]. In addition, the technique was described previously in subsections 2.2.3 and 2.2.4.

Site-specific evolutionary rates were calculated by first calculating posterior probabilities for each site and rate category, using the empirical Bayes approach [70], and then averaging over all rate categories at each site, as described [66].

The FEL model is built on top of the REL model, and can be considered a structure-aware modification of existing FEL approaches [51]. After

determining the best-fitting REL model, we re-fit a GY94 model separately at each site. In this fit, the variables κ and t are not re-fitted to the data but instead set at each site to the values predicted from the REL model. Thus, the FEL approach fits only a single parameter for each site, the evolutionary rate ratio ω.

All data files and analysis scripts are provided as online supplementary materials.

3.3 Results

3.3.1 Estimating site-specific evolutionary rates in a structural context

We constructed separate sequence alignments for human and avian H1, human and avian H3, and human and avian H5 (six separate alignments; see Methods for a detailed description). We then calculated site-specific evolutionary-rate ratios $\omega = dN/dS$ for each alignment. We carried out these calculations in a structure-aware framework; our goal was to determine the relative importance of protein structure and other factors for evolutionary-rate variation among sites.

There are two alternative approaches to calculating per-site evolutionary rates, the random-effects likelihood (REL) method and the fixed-effects likelihood (FEL) method. Under the REL method, a finite set of ω categories is fit jointly to all sites in the alignment. Sites are then probabilistically assigned to categories using an empirical Bayes approach [70, 109]. To obtain distinct

ω values for each site, we calculate averages over all ω categories, weighted by each site's posterior probability to belong to each category [66]. Under the FEL method, a distinct ω value is fit separately to each codon column in the alignment. Existing REL and FEL methods tend to have comparable performance, with either one being preferable in certain applications [51].

We recently introduced a structure-aware REL model, which calculates ω taking into account the relative solvent accessibility (RSA) of individual residues. RSA is a measure of how close to the surface a residue is located in the folded, 3-dimensional protein structure. A residue with an RSA of 0 is located entirely inside the protein structure; it has no contact with water. A residue with an RSA of 1, on the other hand, is completely exposed. Such residues tend to occur in variable surface loops. Intermediate RSA values correspond to partially buried residues.

In the structurally-aware REL model, ω is written as a linear function of relative solvent accessibility (RSA), $\omega = \omega_a \times RSA + \omega_b$, and both ω_a and ω_b are random variables drawn from discrete distributions with a fixed set of categories (see [66] and Methods). The optimal number of slope (ω_a) and intercept (ω_b) categories is determined by minimizing the AIC. This model generally produces a better fit than a comparable model without a structural component. Typical numbers of categories needed for both slopes and intercepts fall between two and four [66].

We fit this model to all six HA alignments, and found that the RSA dependence was significant in all cases except one (human H5). The optimal

number of slopes and intercepts varied by alignment (Table 3.1). We also compared ω estimates predicted by the best REL model with RSA dependence to those predicted by the best REL model without RSA dependence, and found that estimates were generally very similar, with correlation coefficients above 0.9 (Table 3.1). We would like to emphasize, however, that this similarity only arose when we averaged ω over rate categories, as described [66]. Thus, even though using a model with RSA dependence is preferable in terms of model fit, a model without RSA dependence can perform almost as well, as long as ω is averaged over rate categories.

Table 3.1: Comparison of RSA-dependent and RSA-independent model fits.

		slopes, intercepts			
host	HA type	best overall[a]	best no-RSA[b]	best RSA[c]	corr[d]
human	H1	3, 2	0, 3	3, 2	0.95
human	H3	2, 3	0, 3	2, 3	0.98
human	H5	0, 3	0, 3	2, 3	0.99
avian	H1	2, 3	0, 4	2, 3	0.92
avian	H3	2, 2	0, 3	2, 2	0.97
avian	H5	4, 3	0, 4	4, 3	0.96

[a]Number of slopes and intercepts for the overall best fitting model.
[b]Number of slopes and intercepts for the best fitting model without RSA dependence.
[c]Number of slopes and intercepts for the best fitting model with RSA dependence.
[d]Correlation of $\log \omega$ between the best model with and without RSA. All correlations are highly significant ($p < 10^{-10}$).

Next we assessed the reliability of per-site ω ratios, by comparing REL estimates to FEL estimates. Such a comparison has not previously been done

for the structure-aware REL model. We implemented a structure-aware FEL model (Methods), fitted it to all alignments, and found that REL and FEL estimates were generally very similar (Fig. 3.1). Correlation coefficients were typically above 0.9. FEL tended to produce slightly higher ω estimates for large ω values and slightly lower estimates for low ω values. This finding is consistent with the prior observation of shrinkage in REL estimates relative to FEL estimates [51]. One downside of the REL approach is excessive shrinkage if the alignment in small. This problem does not arise in the FEL approach. By contrast, in the FEL approach the number of sequences in the alignment imposes a lower limit on the smallest ω values that can be estimated. All estimates below this limit are simply zero. The REL method, by contrast, assigns rates to these sites based on the density of sites with no non-synonymous variation relative to the density of sites with some non-synonymous variation.

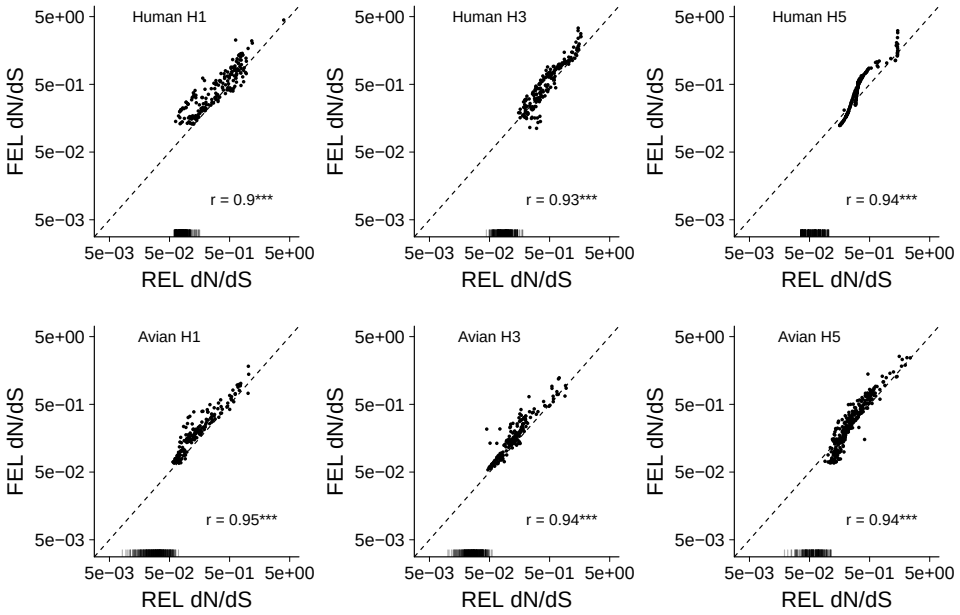

Figure 3.1: Comparison of per-site evolutionary-rate ratios $\omega = dN/dS$ calculated using the REL and the FEL methods. Each dot represents the ω estimate for a single site. Rugs along the x axis indicate sites for which the REL method estimated a positive ω but the FEL method estimated $\omega = 0$. Overall, both methods yield comparable results. Correlation coefficients r are Pearson correlations, calculated for $\log \omega$ excluding sites with $\omega = 0$. FEL tends to estimate slightly higher values for sites with high ω and slightly lower values for sites with low ω. (Significance levels: $^{***}p < 10^{-10}$.)

Because the REL and FEL methods provide comparable estimates and REL provides a positive ω for all sites, we used REL for the remainder of this paper. All results remained qualitatively unchanged when analyses were carried out using the FEL method.

3.3.2 Elevated ω near the sialic-acid binding region

When plotting per-site ω against RSA, we found that ω generally increased with RSA, as expected (Fig. 3.2). However, the rate variation around this overall trend was large. ω varied over one to two orders of magnitude at all RSA values. We found the least rate variation in human and avian H1, and the most variation in human H5.

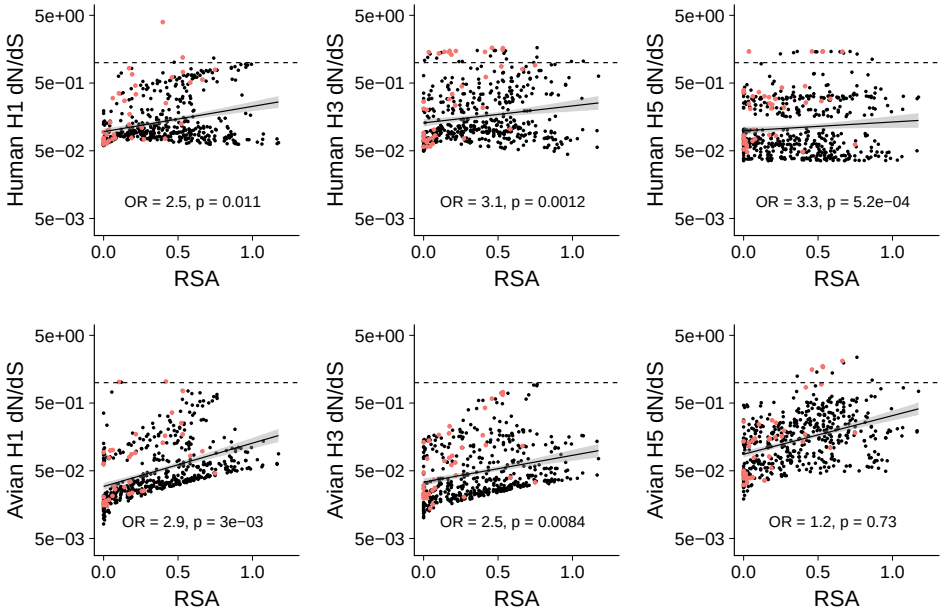

Figure 3.2: Per-site evolutionary-rate ratios $\omega = dN/dS$ plotted against RSA. Each dot represents the ω REL estimate for a single site. Red dots highlight sites in the sialic-acid binding region (SABR sites). Solid lines represent regression lines of ω against RSA, with associated 95% confidence bands. Dashed lines represent $\omega = 1$. OR stands for the odds ratio that sites in the sialic-acid binding region fall above the regression line, with associated p value calculated by Fisher's exact test.

To assess how variation in ω correlated with protein function, we next

compared ω of sites near the sialic-acid binding region to the ω of other sites. We considered sites to be near the sialic-acid binding region if the corresponding residue had at least one atom within 8Å of the sialic-acid molecule in our reference PDB structure. There were 39 such sites. In the following, we will refer to these sites as SABR (Sialic Acid Binding Region) sites. The cutoff of 8Å is conservative, in the sense that we likely captured all sites relevant for sialic-acid binding as well as some sites that are not. Thus, even if the entire sialic-acid binding region was under positive selection, we would not expect all 39 SABR sites to have elevated ω. However, we would expect this set of sites to be enriched for sites with elevated ω relative to other regions in the protein.

Visual inspection of Fig. 3.2 reveals that SABR sites have elevated ω for both human and avian viruses for most HA subtypes. We quantified this association by estimating how much more likely a SABR site was to fall above the ω–RSA trendline than below the trend line. With the exception of avian H5, odds ratios (OR) fell between 2.5 and 3.3 (Fig. 3.2) and were significantly different from 1 (Fisher's exact test). In avian H5, we found no evidence for elevated ω in SABR sites.

3.3.3 Site-specific ω estimates vary substantially among host species and HA subtypes

We next assessed to what extent site-specific evolutionary-rate ratios were comparable across species. We found that while there was a clear trend

towards similar ω across species, variation around this trend was substantial (Fig. 3.3). Correlation coefficients fell between 0.49 and 0.6. Thus, between 24% and 36% of the variation in ω for one host species was explained by the ω values of sequences infecting another host species. More specifically, in any given comparison, strong negative selection (low ω) in one group did not guarantee strong negative selection in another group. Several of the most constrained sites in avian sequences had ω near or above 1 in human sequences (Fig. 3.3).

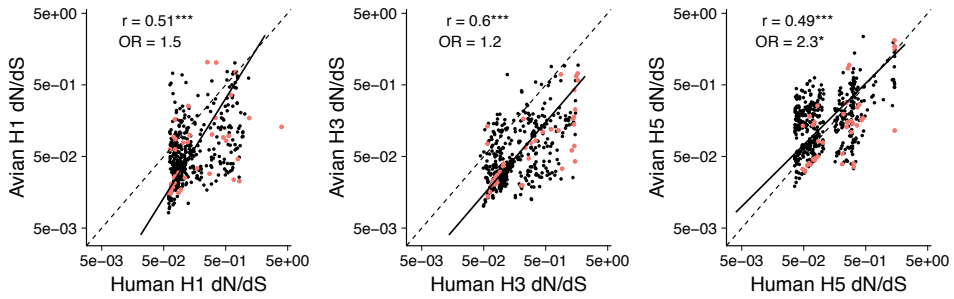

Figure 3.3: Per-site evolutionary-rate ratios $\omega = dN/dS$ plotted for pairs of species. Each dot represents the ω REL estimate for a single site. Red dots highlight sites in the sialic-acid binding region (SABR sites). Dashed lines indicate the $x = y$ line, and solid lines indicate the direction of maximum covariation in each data set. The correlation coefficients r represent Pearson's correlations, calculated for $\log \omega$. OR stands for the odds ratio that sites in the sialic-acid binding region fall below the solid line, with associated p value calculated by Fisher's exact test. (Significance levels: $^{*}p < 0.05$; $^{***}p < 10^{-4}$.)

Despite the overall finding that conservation in one group does not guarantee conservation in the other group, most comparisons also showed clusters of sites with very similar (and typically low) ω values. Thus, there seems to

be a core set of conserved sites that are shared among strains.

We found systematic differences in H1 and H3 across host species; ω estimates for avian H1 and H3 were generally lower than the corresponding estimates for human H1 and H3 (Fig. 3.3). Human and avian H5, on the other hand, showed virtually no difference in ω on average.

For SABR sites specifically, we found that their ω differences were mostly in agreement with the overall ω differences of the strains compared. We tested whether SABR sites were more likely to fall either above or below the lines of maximum covariation in Fig. 3.3, using Fisher's exact test as before. For H1 and H3, we could not reject the null hypothesis of an odds ratio of 1. However, we did find that SABR sites evolved at a significantly elevated ω in human H5 compared to avian H5.

3.3.4 Differences in ω are moderately biased towards the protein core

Finally, we wanted to assess whether the largest differences in ω among groups associated with specific regions in the structure. To this end, we introduced a function G, defined as $G(1,2) = |\log(\omega_1/\omega_2)|$, which measures the absolute difference in $\log\omega$ at a site. We plotted G as a function of RSA to determine whether the largest differences occurred on the protein surface or in the core (Fig. 3.4).

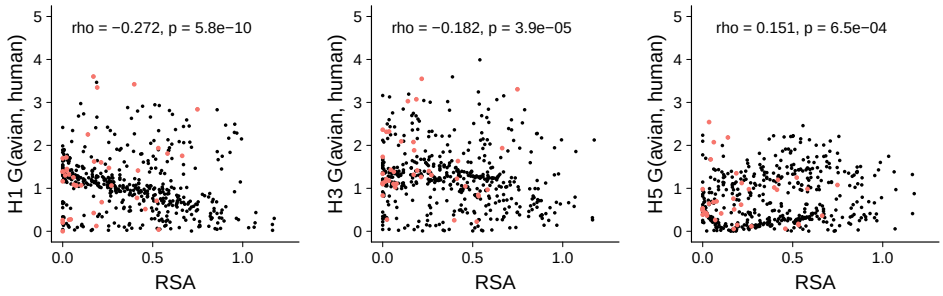

Figure 3.4: Cross-species rate differences as a function of RSA. The function $G(1,2)$, defined as $G(1,2) = |\log(\omega_1/\omega_2)|$, is plotted against RSA. Correlation coefficients are Spearman correlations between G and RSA. The function G is negatively correlated with RSA for H1 and H3, and positively correlated for H5. Red dots highlight sites in the sialic-acid binding region (SABR sites).

We found a weak trend of declining G with increasing RSA in H1 and H3, and the opposite in H5 (Fig. 3.4). Thus, for H1 and H3, sites in the core of the protein showed, on average, slightly larger differences in ω than sites near the surface. This finding suggests that ω differences among strains are not just caused by different antigenic pressure (which would exert itself on the surface of the protein) but also by some processes that act throughout, and possibly particularly towards the core of, the protein structure. The inverted trend for H5 may indicate that in these viruses, the difference between the avian and the human selective environment is indeed dominated by surface-bound processes (antibody binding and/or sialic-acid binding).

We mapped the SABR sites onto the G–RSA plots to determine whether they were contributing to an excess of buried sites with high ω in humans. Clearly, many of the SABR sites have relatively low RSA (Figs. 3.2 and 3.4).

However, no obvious trend emerged. SABR sites did not cause the negative correlation between G and RSA observed for H1 and H3.

3.4 Discussion

We have found that protein structure makes a significant contribution to per-site evolutionary rate variation in influenza hemagglutinin. We have also found that sites near the sialic-acid binding region show elevated ω in most strains. Finally, we have found that site-specific rate variation among different influenza strains is comparable but by no means identical. This finding emphasizes the diverse nature of strain and host-specific selection pressures throughout the HA protein.

The strength of the correlations among ω for different host species inform us about the extent to which protein structure constrains sequence evolution. If protein structure were entirely responsible for any rate variation, then these correlations should be near 100%. By contrast, if protein structure had no influence on sequence variation, then these correlations should be near 0% (to the extent that evolution has truly been restricted to only one host species). We found that the ω values of sequences for one host species explained between 24% and 36% of the variation in ω for sequences of another host species (correlation coefficients fell between 0.49 and 0.6). Thus, at least 60%, and possibly more, of the variation in ω is not explained by protein structure.

We found that all viral strains except avian H5 experienced elevated ω in the sialic-acid binding region. This finding likely reflects diversifying selec-

tion in this region, driven by antibody escape or HA–receptor interactions. In humans, antibodies are not generally thought to bind at the sialic-acid binding region, yet such antibodies do exist and may be sufficiently common to drive positive selection in this region [103]. Further, the sialic-acid binding region is small compared to the typical binding region of an antibody. Thus, mutations in the periphery of the receptor pocket can interfere with antibody binding without inhibiting sialic-acid binding [103]. Second, the positively selected mutants may not primarily be antibody-escape mutants but rather mutants that change the avidity of HA–receptor interactions [41]. Such substitutions could be driven by sequential passage in immune and naive hosts, for example human adults and human children, where immune hosts select for increased avidity and naive hosts select for decreased avidity [41].

In our comparison of strains infecting different host species (Figure 3.3), we found for H1 and H3 that evolutionary-rate ratios ω in the sialic-acid binding region were consistent with the overall differences in evolutionary rates among host type. For example, the comparison of avian to human H3 shows that avian H3 is overall more conserved than human H3, and this increased conservation is visible in similar magnitude near the sialic-acid binding region and elsewhere in the protein. Further, both the H1 and the H3 comparison show clear clusters of sites that are conserved in both strains but overall more conserved in one strain than in the other (i.e., clusters are shifted away from the dashed line representing equal ω). Note that this result is not likely to be caused by mutation-rate differences, since $\omega = dN/dS$ is normalized for

mutation rate. In combination, these findings suggests that there are systematic differences in selection pressures among human and avian influenza strains. These altered selection pressures seem to operate throughout the entire protein, not just in specific regions. Previous studies have shown that the sialic-acid binding region is more conserved in avian viruses than in human viruses [63]. Our results do not disagree with this previous observation, but they suggest that this observation may have been caused by selection pressures acting on the entire HA protein, rather than by selection pressures acting specifically on the antigenic regions of the protein.

It was surprising that human and avian H5 sequences showed substantial differences in their evolutionary patterns, and in particular, that sites in the sialic-acid binding region showed elevated ω in human H5 compared to avian H5. In contrast to H1 and H3, human and avian H5 sequences cannot be considered distinct viral lineages. Because H5 cannot be transmitted directly among humans, one would expect that human H5 sequences should look like a random sample of avian H5 sequences. Yet this was not the case. In particular, sites in the sialic-acid binding region evolved significantly differently in human and avian H5. There are two possible explanations for this observation: First, the human physiology starts exerting a selection pressure on the sialic-acid binding region as soon as a human is infected with H5, and the H5 sequences evolve significantly within patients. Second, and more likely, only a subset of H5 sequences, in particular sequences with specific changes in the sialic-acid binding region, can infect humans. Therefore, human H5

sequences are not a random sample from avian H5 sequences. Prior work has shown the accumulation of key adaptive substitutions after introduction of avian H1, H2, and H3 to mammalian hosts [62]. The methods we used here are not suitable to identify key substitutions in H5 adaptation from avian to human hosts; however, suitable methods exist [54] and could be applied to this problem in future studies.

A number of prior studies have looked at adaptive evolution in influenza HA, using both dN/dS-based methods (e.g. [15, 94, 106]) and other methods (e.g. [8, 54, 96]). However, a site-by-site comparison across different host species, as we have performed here, is not usually done. One exception is the recent work by Tamuri et al. [96], who identified sites under differential selection pressure in human and avian influenza. Their results are broadly consistent with ours. The majority of sites showed no significantly different selective constraints among human and avian sequences. Nevertheless, a number of sites did show significant differences. For H3, Tamuri et al. [96] identified 11 such sites. This number seems low relative to the large amount of ω differences between human and avian H3 we found here. Several issues may contribute to this apparent discrepancy: (i) only 11 sites were identified *with high confidence* by Tamuri et al., more sites were identified at lower confidence; (ii) we did not attempt to identify for which sites ω was significantly different among species here; while we saw a substantial amount of variation in ω, some of this variation may not be statistically significant; (iii) Tamuri et al.'s and our model are fundamentally different and may identify different sites.

To assess whether sites in the sialic-acid binding region had elevated ω, we used Fisher's exact test instead of a seemingly more natural linear regression of ω against RSA and site type. We decided against the regression analysis because ω is a derived quantity, estimated from a model that used RSA as input data. Therefore, the regression analysis could potentially yield incorrect results. By contrast, for any chosen bisection of the data, Fisher's exact test answers the question whether SABR sites are enriched on one side of the bisection. Thus, even if the trendlines in Fig. 3.2, derived from a regression, are not entirely correct, we can conclude that SABR sites are significantly enriched above the trendlines for all strains except avian H5. One caveat of this application of Fisher's exact test is that we are treating ω values as observations, even though they are noisy estimates. The noise on ω reduces the power of Fisher's exact test; however, we do not expect the noise to artificially inflate the rate of false positives.

We fitted both RSA-dependent and RSA-independent models to all alignments, and found that in all except one case the RSA-dependent model produced the better fit. Due to the way we chose to weight per-site evolutionary-rate estimates by each site's posterior probabilities, in all cases ω estimates obtained by RSA-independent models were highly correlated with ω estimates obtained by RSA-dependent models (Table 3.1). Hence, rather than explicitly defining an RSA-dependent model, one could make estimates within an RSA-independent framework; then, structural information could be inferred in post-analysis by taking into account partial occupancy of multiple rate categories

for each site in the protein. Our analysis shows that incorporating structure in this way produces less accurate rate estimates than when the structure is already known. However, for future studies, this finding also suggests that even when analysis in a structural context is desired, it may be sufficient to first calculate evolutionary rates without structural input and then correlate these rates with structural features. While rates calculated with structural input are expected to be more accurate, and can reveal aspects of the data that might have remained obscure otherwise [66], there are likely many applications where the saved computational expense of not fitting structure-aware models is worth the small loss in the accuracy of estimates.

Throughout our analysis, we assumed that the protein structure (and hence RSA) remains constant over time and even across HA subtypes. This approximation is reasonable, because hemagglutinin is a structurally homogeneous protein; available crystal structures show little structural variation. The full-length form of hemagglutinin, referred to as HA0, consists of over 500 amino acids. Prior to capsid incorporation and viral burst, HA0 is cleaved into two proteins, HA1 and HA2. Comparing all structures available in the protein database (over 100, both pre- and post-cleavage), we found that greater than 90% of the two post-cleavage subunits varied by less than 2 Å RMSD (not shown). Moreover, most of that variation was the result of flexible regions near chain terminations. Several post-cleavage structures exist; they tend to have very different structures near the inter-chain break site of HA0, due to the additional chain terminations. To eliminate the unpredictable variation

near the site of inter-chain termination, we chose to base our RSA calculations on an HA0 crystal structure. Since the cleavage site is in the transmembrane domain of HA, approximately 80 Å away from the sialic-acid binding region, we do not expect any structural variation in this region to impact our results in any substantive way. Finally, we used a high quality structure from H1 influenza throughout this work, because there were no suitable, full-length H3 or H5 HA0 structures.

The high evolutionary conservation of hemagglutinin structures across subtypes and host-species implies that the immune systems of humans and birds do not generally rely on the 3-dimensional structure of hemagglutinin for binding, but only on the specific sequence of amino acids that is overlayed on that structure. If structure itself were important for immunogenicity, we would expect to see large structural deviations as the host immune system pressured influenza to escape neutralization.

Chapter 4

Evaluating the effect of mutations in a protein–protein interaction

4.1 Introduction

This work was published previously in the journal *PeerJ*.[1]The computational prediction of mutational effects on protein–protein interactions remains a challenging problem. Several methods are available to perform an energy difference calculation from an experimentally determined co-crystal structure. For example, end point methods can be performed rapidly, with relatively low computational cost [37, 50]. However, such methods can suffer from various simplifying assumptions. For example, they generally use an implicit solvent approximation and assume the end state difference with minimal structural rearrangement is sufficient to discriminate energetic differences [37, 50]. Alternative approaches have been developed using machine learning, training coefficients in a weighted equation containing geometric and energetic parameters [4, 44, 100, 101]. Unfortunately, such machine-learning approaches often suffer in novel applications, for which available training sets are small

[1]A. G. Meyer, S. L. Sawyer, A. D. Ellington, and C. O. Wilke. Analyzing machupo virus-receptor binding by molecular dynamics simulations. PeerJ, 2:e266, 2014. S. L. Sawyer, A. D. Ellington, and C. O. Wilke helped to design the project and write the manuscript.

or non-existent. As such, these methods are poorly suited for most host-virus protein–protein systems. By contrast, first principles methods can forgo training, but currently available methods such as free energy perturbation (FEP) and thermodynamic integration (TI) rely on a transitional model (where one state may be wild-type and the other may be a mutant) to make rigorous free energy calculations [18, 31, 40, 61]. While these may be considered two of the gold standard techniques for calculating affinity differences, there are a huge number of theoretical and technical complexities that must all be properly managed to ensure a converged solution [39]. Such considerations quickly come to dominate the protocol, and the necessary book keeping introduces the possibility of human error [39]. Moreover, as the two ending states look ever more dissimilar the chances of convergence fall rapidly. To ensure convergence, these techniques are typically limited to small differences (such as point mutant comparisons) with a few, very impressive exceptions [39, 40, 102]. For most investigators, larger differences quickly become intractable as the number of intermediate steps required to compute a converged solution grows or the complexity of adding restraining potentials and computing approximations expands [39, 40, 102].

Here we propose that much of these complexities can be avoided if all we are interested in is a relative comparison of the effects of different mutations on protein-protein interactions, rather than measuring an absolute or relative binding affinity with experimentally realistic units. We impart a pulling force within an all-atom molecular dynamics simulation on one member of the com-

plex while the other is held in place. Then, we measure the force required for dissociation [38, 45, 46, 60, 71, 73]. Although such biasing techniques are commonly used in protein-ligand binding problems, they are less commonly applied to protein–protein interactions, and almost never to mutational analysis in a protein–protein system. This is largely the result of free energy convergence difficulties and computational limitations [22, 23]. Using a proxy for relative binding affinity rather than calculating absolute affinities can solve these problems. Here, as proxies, we use the maximum applied force required for separation and the area under the force-versus-distance curve (AUC). For comparison, we also calculate relative free energy differences using the traditional dual topology FEP paradigm, and we show that the two approaches yield congruent results.

We used SMD and FEP to interrogate the interaction between machupo virus (MACV) spike glycoprotein (GP1) and the human transferrin receptor (hTfR1) [1, 17]. Machupo virus is an ambisense RNA virus of the arenavirus family [17]. Worldwide, arenaviruses represent a significant source of emerging zoonotic diseases for the human population [17]. Members of the arenavirus family include the Lassa fever virus endemic to West Africa, the lymphochoriomeningitis virus (LCMV) endemic to rodents in several areas of the United States, and the Guanarito, Junin, and Machupo viruses endemic to rodents in South America [17]. The South American arenaviruses typically infect humans after rodent contamination and can cause a devastating hemorrhagic fever with high mortality [17].

The hTfR1 is the primary receptor used by MACV for binding its host cell prior to infection. The primary role of hTfR1 *in vivo* is to bind transferrin for cellular iron uptake. The hTfR1 protein contains three extracellular domains: two basilar domains and an apical domain. The two basilar domains serve most of the transferrin-binding function [1,79]. Viral entry is initiated by GP1 binding to the apical domain of hTfR1. Previous work has indicated that the GP1/hTfR1 binding interaction is the primary determinant of MACV host range variation [19, 79]. The co-crystal structure shows that the high affinity interaction between GP1 and hTfR1 forces the normally flexible loop in the apical domain of hTfR1 into a rigid β-pleated sheet domain. For GP1, several extended loops mediate binding to hTfR1 [1, 79], and many of the interface interactions are mediated by extensive hydrogen-bonding networks [1]. Experimental alanine-scanning and whole-cell infectivity assays have identified several sites in both GP1 and hTfR1 that are probably critical for establishing infection [19, 79].

We applied our computational method to wild type (WT) and mutant complexes, and found that we could resolve relative differences in unbinding and predict significant affinity changes. Importantly, the affinity changes predicted using only max force or AUC show a strong correlation with rigorous relative free energy differences computed by FEP. At sites known to be important for successful viral entry, we found that the biochemical cause of reduced infectivity may not be as simple as the static structure suggests. For example, the static structure shows a hydrogen-bonding network connected to site N348

in hTfR1. According to our simulations, this network may not affect binding affinity directly. In addition, our study offers an all-atom steered molecular dynamic approach to avoid some of the pitfalls of several existing methods used to evaluate mutations in protein–protein interfaces.

4.2 Materials and Methods

4.2.1 System Modeling

For our experiments, we used the experimentally determined GP1/hTfR1 structure (PDB-ID: 3KAS) [1]. The apical domain of hTfR1 interacts directly with GP1 while the other two domains are closer to the cell membrane and have essentially no interaction with GP1. The biophysical independence of the apical domain allowed us to isolate it without significantly affecting the GP1/hTfR1 interaction.

We used the protein visualization software PyMOL [87] to remove residues 121-190, 301-329, and 383-756 in the hTfR1. No residues were removed from the viral protein. Figure 4.1 shows a model of the initial structure and that of the pared structure. Although GP1 has several glycosylatable residues, we opted to use the de-glycosylated protein for this study. The complexity of correctly parameterizing diverse sugar moieties is outside of the scope of this paper. Furthermore, although it is known that GP1 is glycosylated, and some of those sugars contact hTfR1, the sugars in the available PDB structure are not physiological for mammals [1]. In total we removed 10 sugars from the crystal structure for this study.

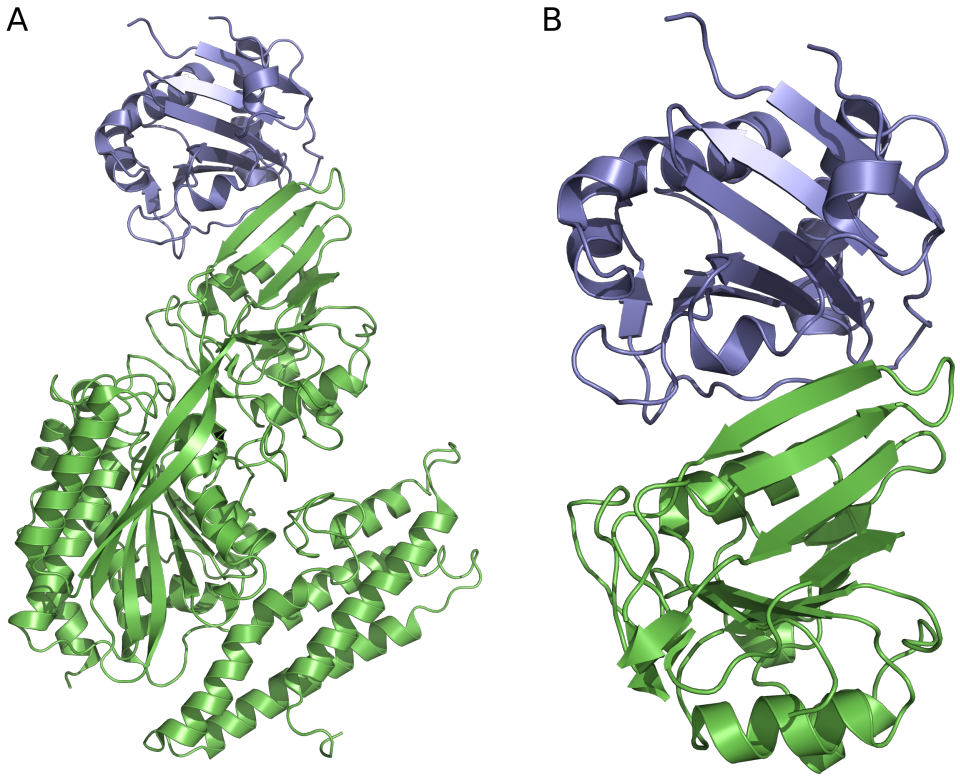

Figure 4.1: The GP1/hTfR1 complex. GP1 is shown in blue and hTfR1 is shown in green. (A) The full, de-glycosylated GP1/hTfR1 co-crystal structure. (B) The reduced structure used in SMD simulations.

After system reduction, the Visual Molecular Dynamics (VMD) [43] package along with its system of back-ends was used for all subsequent modeling. The Orient add-on package allowed us to rotate the system axis such that the direction of steering was oriented directly down the z-axis. Deglycosylation simplified the system such that Autopsf could easily find the chain terminations and patch them appropriately. The Solvate package was used to generate a TIP3P water model with a 5 Ångstrom buffer (relative to

the maximum dimensions of the proteins) on all sides except down the positive z-axis where a 20 Ångstrom buffer was created. Finally, we used the Autoionize package to place 150 millimolar NaCl and neutralize the total system charge. In the end, each modeled system had approximately 28,000 atoms.

4.2.2 Equilibration

NAMD was used for all simulations in this study [75]. In addition to the modeled system, for equilibration we generated a configuration file that fixed the α-carbon backbone. This was accomplished by setting the B-factor column to 1 for the fixed atoms and to zero for all other atoms. Further, we generated a configuration file with fixed α-carbon atoms at residues 41-92 (numbered linearly, in this case, starting at 1 for the first amino acid as was required for NAMD) in the hTfR1. The second file was used to affix a harmonic restraint, thus preventing any unfolding due to system reduction. More importantly, the harmonic restraint allowed the protein complex to equilibrate while preventing any drift from its predefined position; the restraint did not constrain the structure of each protein, or the relative position or orientation of the two proteins to each other. Finally, we calculated the system center and dimensions for use in molecular dynamics settings. The exact NAMD configuration files are available on github (https://github.com/clauswilke/MACV_SMD).

We used the Charmm27 [12] all-atom force field. The initial system temperature was set to 310K. Several typical MD settings were used including

switching and cutoff distances (see provided configuration files). In addition, we used a 2 femtosecond time step with rigid bonds. We used periodic boundary conditions with the particle mesh ewald (PME) method of computing full system electrostatics outside of the explicit box. Furthermore, we used a group pressure cell, flexible box, langevin barostat, and lavegin thermostat during equilibration. A harmonic restraint (called harmonic constraint in VMD) was set as stated previously.

To start the simulation, the barostat was switched off and the system was minimized for 1000 steps. Next, the fixed backbone was released, and the system was minimized for an additional 1000 time steps. Subsequently, the system was released into all-atom molecular dynamics for 3000 steps. Finally, the langevin barostat was turned on and the system was simulated for 2 ns (1,000,000 steps) of chemical time. For each mutant, twenty independent equilibration replicates were run with an identical protocol.

4.2.3 Steered Molecular Dynamics

We used the final state from each equilibrated system to restart another MD simulation. Our steering protocol is fundamentally similar to [22] with slightly different parameter choices. Perhaps the one significant difference lies in our choosing to not use a thermostat or barostat. We can make this choice because we are not trying to calculate the binding free energy by any physically rigorous approach (the Jarzynski inequality being one example). Following equilibration, the final state of each simulation was used to generate

a configuration file fixing the α-carbon on residues 1, 58, 73-83, 96, 136, 137, 138, and 161 (again with linear numbering) in the hTfR1. These residues were selected as they are far from the binding interface and sufficiently distributed to prevent any orientational motion of the receptor relative to the viral spike protein. The center of mass of the α-carbons of all residues (163-318 in linear numbering) in GP1 received an applied force during the simulation. The NAMD convention does not actually apply a force to all α-carbon atoms but rather uses the selection to compute an initial center of mass. Then, during the steering run, the single center of mass point is pulled with the parameters described below. We used the same force field parameters (exclude, cutoff, switching, etc.), the same integrator parameters (time step, rigidbonds on, all molecular being wrapped, etc.), and the same particle mesh ewald parameters as in equilibration. Periodic boundary conditions were incorporated as part of the system (as is the convention in NAMD restart) and PME was again used to approximate full system electrostatics.

We ran test simulations at several force constants and visually inspected the results. A force constant of 5 kcal/mol/Å^2 was chosen due to its relatively low signal-to-noise ratio. This constant is slightly lower than the more common 7 kcal/mol/Å^2 found in several recent studies; that value is commonly selected primarily because it is the force constant found in the SMD tutorial available through the NAMD developers. Moreover, the force constant could very likely be set to a range of nearby values with little loss in predictive power.

In SMD experiments the pulling velocity should be as low as possible for

the available computational time [22,23]. We choose a velocity of 0.000001 Å/fs = 1 Å/ns, and direction down the positive z-axis. One could use faster pulling if the computing time must be reduced, but slower than necessary pulling speeds are not typically considered problematic.

SMD was run for 15 ns (7,500,000 time steps) of chemical time. For each simulation, we randomly selected one of the equilibration runs for restart. We ran 50 replicate simulations per mutant for a total of 550 SMD simulations. All GP1/hTfR1 complexes separated by greater than 4 Å and many separated to 10 or more.

To leave the final trajectory of a tractable size, only 1000 evenly spaced frames were retained from each simulation, leaving a final trajectory size of 323 MB. See the supplemental movie for a representative unbinding trajectory. Initial development of the SMD protocol was carried out on the Lonestar cluster at the Texas Advanced Computing Center (TACC). All production SMD simulations were performed on the Hrothgar cluster at Texas Tech University, using NAMD 2.9. Each simulation was parallelized over 60 computational cores and utilized approximately 20 hours of computing time. The total chemical time simulated for this project was nearly 10 μs, requiring slightly over 1 million cpu-hours.

4.2.4 Free Energy Perturbation

Briefly, we used the traditional dual topology approach to FEP [30,74]. This involves a thermodynamic cycle where a set of atoms are progressively

decoupled from the environment while another set of atoms are progressively coupled. To compute the relative free energy difference requires knowing the free energy change when the transformation is carried out for the bound complex and the individual protein. Then, one can compute the relative free energy difference between a WT and mutant complex by taking the difference between the energy required to decouple/couple the atoms in solution from the energy required to decouple/couple the atoms in the bound complex [30, 74].

Again, the NAMD configuration file is made available via github (https://github.com/clauswilke/MACV_SMD). We used a similar configuration to that in equilibration. One significant difference was to make a cubic water box with a side length equal to the long axis of the complex plus a 10 Å buffer on either side, and simply restrict center of mass motion with the NAMD setting. This was done to avoid affecting the system energy while calculating free energy differences.

The transition protocol for bound and free protein systems were identical. They started with 1000 steps of minimization and 250,000 steps of equilibration in the starting state for the forward and reverse directions. Phase transitions were carried out in steps of λ=0.05. Each transition was carried out for 250,000 steps. The first 100,000 steps after phase transition were reserved for equilibration and the final 150,000 steps were used for data collection.

The VMD mutator tool was used to generate the necessary topology file and the parseFEP tool [59] in VMD was used for subsequent analysis. We used it to perform error analysis and compute the Bennett acceptance ratio as

the maximum likelihood free energy difference of the two states under consideration. Though the larger transitions presented difficulty in a small number of windows, forward and reverse hysteresis was generally in good agreement for all complexes. The double mutants were performed by first doing the Y211A mutation followed by the other of the two mutants. Then, the ΔG's were simply added together to get the total energetic difference.

4.2.5 Post-processing

The python packages MDAnalysis [67] and ProDy [5] were both used at various points in post-processing. The molecular trajectory (comprising the atomic coordinates per time) was parsed to compute the center-of-mass for each of the two complexes. The starting center-of-mass distance was set to zero and the distance was re-computed at each time step relative to the starting distance.

The statistical package R was used for all further analysis and visualization. Each of the 50 independent trajectories per mutant produced a fairly noisy force curve. The force curves for each mutant were smoothed over all replicates by using the smooth.spline() and predict() functions in R with default settings. The two primary descriptive statistics we used were maximum interpolated applied force and total area under the interpolated curve (AUC). We tested for signifiant differences in maximum force or AUC by carrying out t tests for all pairwise combinations (each mutant compared to each other mutant), using the pairwise.t.test() function in R. We adjusted p values to

correct for multiple testing using the False-Discovery-Rate (FDR) method [7]. The ggplot [104] package was used to generate most of the figures.

Analysis scripts and final data (except MD trajectories) are available on the github repository accompanying this publication (https://github.com/clauswilke/MACV_SMD).

4.3 Results

4.3.1 The GP1/hTfR1 system

The GP1/hTfR1 interface (Figure 4.2) marks a particularly important and useful test system. There are several sites on both the human and viral protein known to affect the infectivity phenotype of MACV. Many of the important sites have been mapped by *in vitro* flow-cytometry based entry assays. The GP1/hTfR1 interface appears not to be dominated by one particular type of interaction (electrostatics, hydrogen-bonding, or van der Waals). In addition, much of the binding domain on hTfR1 is on a loop that is flexible prior to viral binding, but organizes to become a strand of a β-sheet on binding. As a result, many other computational techniques [37, 50] are only marginally useful. The complex nature of this interface represents a particularly difficult challenge for traditional computational analysis.

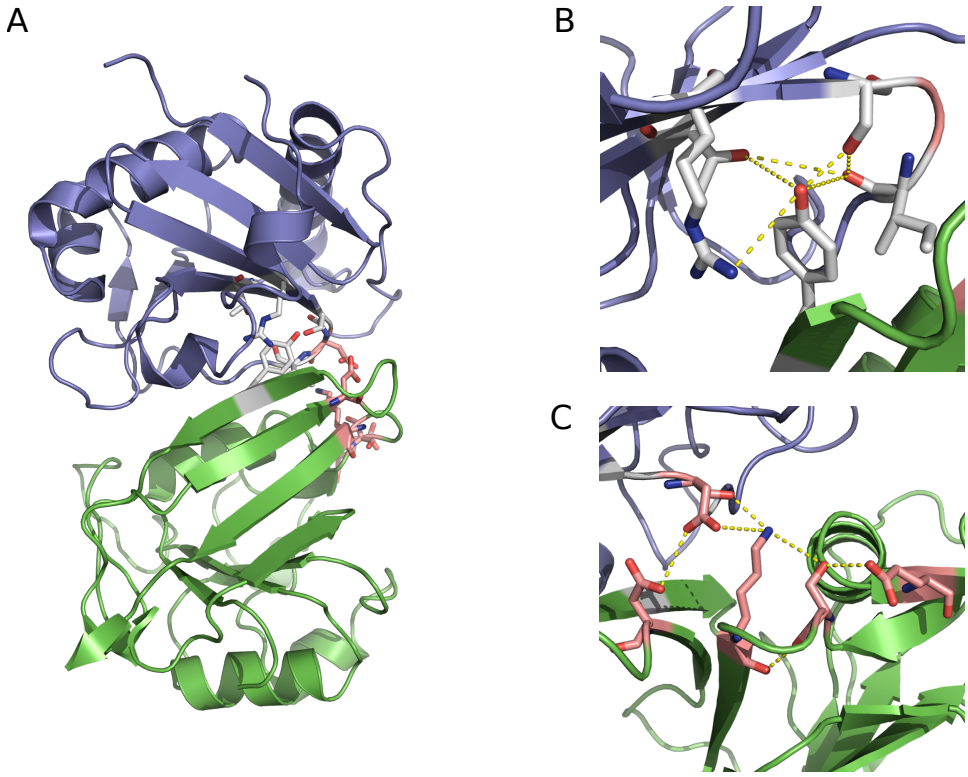

Figure 4.2: The two hydrogen bonding networks. GP1 is shown in blue and hTfR1 is shown in green. (A) The first network including Y211 and R111 is shown in white, and the second network containing N348 is shown in pink. (B) Near view of the first network with contacts in yellow. (C) Near view of the second network with contacts in yellow.

In total, we tested 7 point mutants and 3 double mutants in addition to the WT complex (Table 4.1). All of the mutations are within 5 Å of the protein–protein interface. Mutations in hTfR1 at site 211 have proven capable of causing loss-of-entry according to *in vitro* flow-cytometry infection assays or known host-range limitations [19,78,79]. Most likely, this effect is caused by the destruction of a critical hydrogen bond to Ser113 or Ser111 in GP1. The

lost hydrogen bond would lead to the subsequent loss of a large hydrogen-bonding network seen in the crystal structure (Table 4.1) [1]. In a manner similar to site 211, N348 appears to be important for binding by participating in a critical hydrogen bonding network [1, 78] to GP1. In particular, N348Lys is reported in the literature to cause significantly reduced viral entry *in vivo* (Table 4.1) [1, 78]. Finally, an alanine mutation at site 111 in GP1 (mutation vR111A) has also been shown to cause decreased entry (Table 4.1) [79]. For notation purposes, the viral site is always referred to with a preceding 'v'.

Table 4.1: Summary of prior information available for each mutation tested. Observed *in vivo* refers to mutations that have been observed in rodent populations. Phenotype *in vitro* refers to the observed phenotype in *in vitro* viral entry assays.

Mutation	Observed *in vivo*	Phenotype *in vitro*
WT	Yes	Normal Entry
N348A	No	-
N348K	Yes	Diminished Entry
N348W	No	-
vR111A	No	Diminished Entry
N348A/Y211A	No	-
vR111A/Y211A	No	-
Y211D	Yes	No Expression
Y211T	No	Diminished Entry
Y211A	No	No Expression
N348W/Y211A	No	-

Despite the fact that viral binding occurs at the site of a flexible loop in the free hTfR structure, our data shows after binding the strand is extremely

rigid. In the bound conformation, only two sites of the loop have root mean squared fluctuation (RMSF) values in the top half of all receptor sites during equilibration (Figure 4.3), and those are almost completely exposed to solvent. This is unsurprising considering the high degree of burial that occurs as a result of viral binding. Computing the root mean squared deviation (RMSD) of the entire structure over the trajectory shows that none of the mutations are so deleterious as to cause rapid unbinding. In fact, the RMSD over trajectory looks highly invariant across mutants (Figure 4.4). In the unbound state, calculated near the end of the SMD trajectory, all of the residues in the WT receptor interfacial strand are in the top half of RMSF over all receptor sites (Figure 4.5). Thus, if sufficient simulation time is not dedicated to allowing this unfolding process, standard free energy techniques may miss the energetic contributions that result from ordering the flexible loop in the hTfR apical domain.

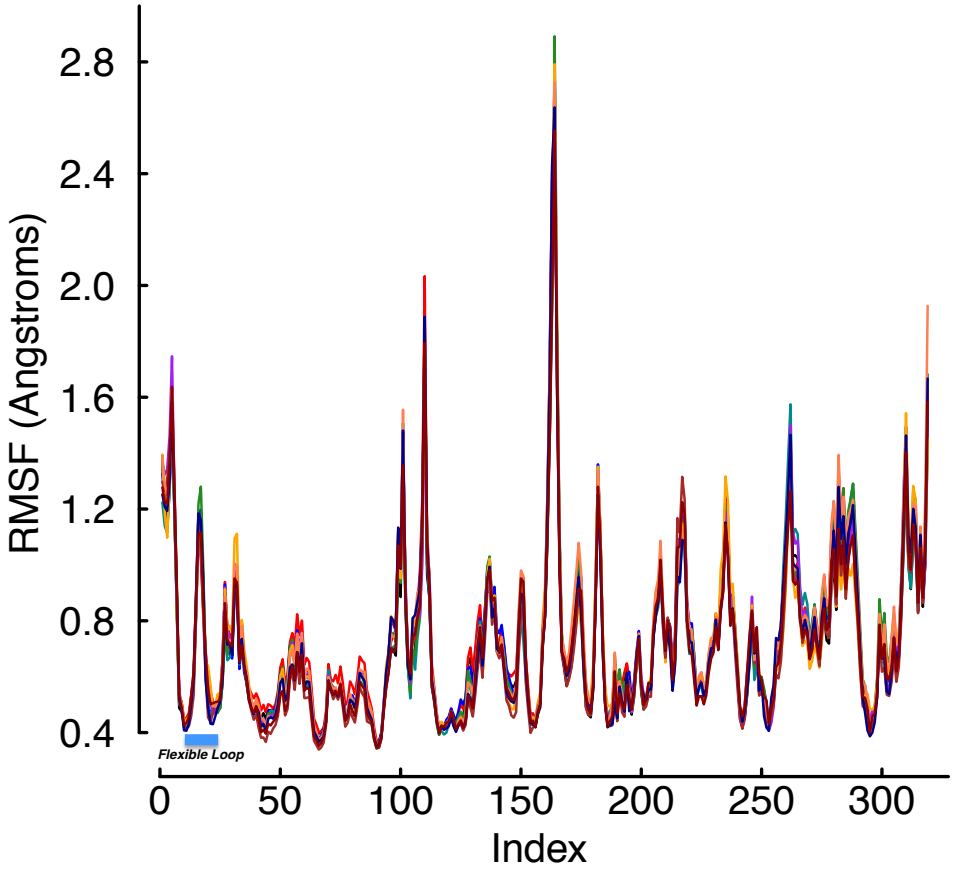

Figure 4.3: RMSF values during equilibration. The RMSF values for every site in the bound complex computed during the equilibration phase of the protocol. Each color represents the average over 20 trajectories of a single mutant. Indices 17-25 are the hTfR flexible loop. The plot shows the flexibility of each site is essentially independent of mutation, and two sites (indices 17 and 18) above 0.72 Å are a part of the flexible loop in the free receptor. However, these two residues are not actually found in the protein–protein interface, but rather are almost completely solvent exposed with the virus bound.

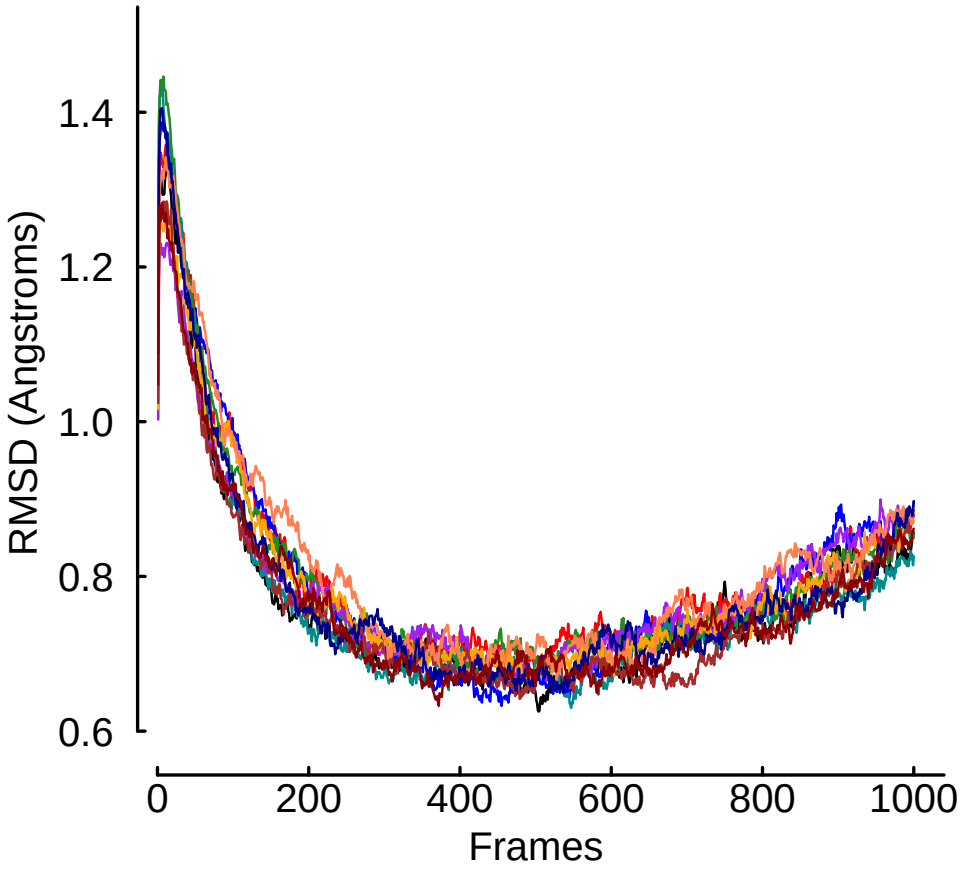

Figure 4.4: RMSD values during equilibration. The RMSD values over the time of the trajectory computed during the equilibration phase of the protocol. Each color represents the average over 20 trajectories of a single mutant. The plot shows none of the mutants causes immediate unbinding of the protein–protein complex. In addition, the universal upward trend near the end of the equilibration trajectories may indicate the crystal is more tightly packed than would normally occur in solution.

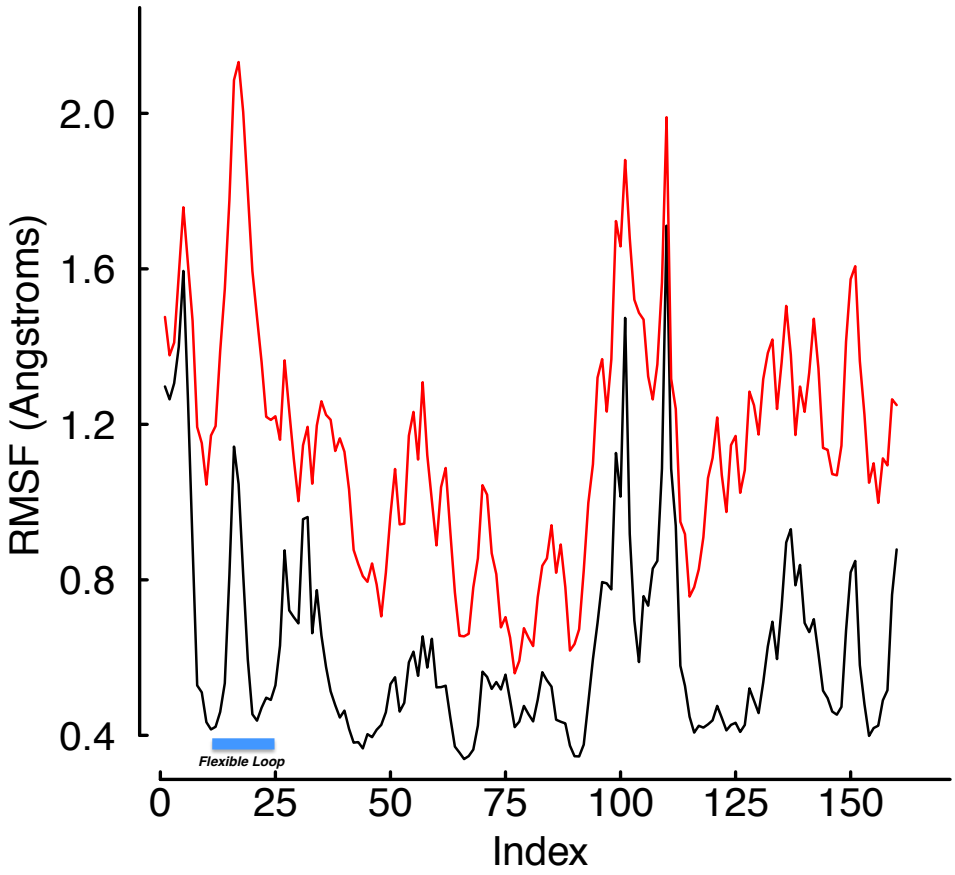

Figure 4.5: RMSF values of WT hTfR in equilibration and SMD. The RMSF values for every site in the WT receptor were computed during the equilibration phase and during final 50 frames of the SMD trajectories. The black line was computed over equilibration and the red line during SMD. The plot shows the solution mobility of the hTfR flexible loop increases more than the average during the unbinding process.

4.3.2 Molecular dynamics simulations

We analyzed the GP1/hTfR1 system using two molecular dynamics techniques. First, by carrying out SMD using a known force constant and

pulling with a constant velocity, we could calculate the applied force during protein–protein dissociation [22,23]. A typical averaged force curve comparison can be seen in Figure 4.6, and individual images of all averaged force curves are available in the associated github repository, in folder figures/force_curves. As seen in Figure 4.6, the dissociation distance was relatively consistent among mutants. The supplementary movie visually illustrates the separation distance between peptide domains. The quantities maximum applied force and AUC were derived from the force-versus-distances curves. Their summary statistics are reported in Table 4.2. As we are more interested in the phenotypic impact of interface mutations we avoided many of the more physically rigorous, but technically complicated calculations that are possible with SMD [45, 46].

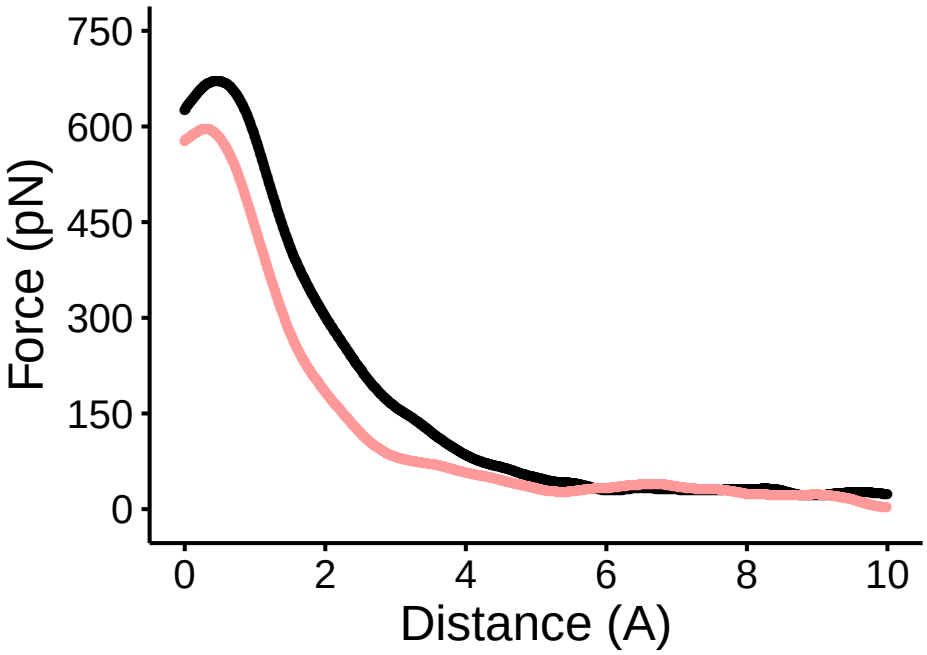

Figure 4.6: Force versus distance curve of WT and the Y211A mutant. The average force curve for 50 replicates of the WT complex is shown in black, and the average of 50 replicates of the Y211A mutant is shown in red. There is a large difference in both maximum applied force and AUC between the two complexes.

Table 4.2: Summary statistics for each mutation tested. μ_{MAF} is the mean in piconewtons and σ_{MAF} is the standard deviation of maximum applied force over all simulations. μ_{AUC} is the mean and σ_{AUC} is the standard deviation of AUC over all simulations. ΔG is the free energy difference in kcal/mol calculated via FEP by the dual topology paradigm.

Mutation	μ_{MAF} (pN)	σ_{MAF}	μ_{AUC}	σ_{AUC}	ΔG (kcal/mol)
WT	734.4856	131.6513	145460.4	60232.26	0.000
N348A	748.5217	137.4864	133913.9	51078.64	-2.149
N348K	705.0707	108.5079	141084.4	54450.28	+3.184
N348W	697.3642	132.6436	136886.0	53796.44	+3.033
vR111A	713.8081	106.7374	136103.2	52070.85	+0.466
N348A/Y211A	703.7027	128.5866	113464.2	57451.62	+5.203
vR111A/Y211A	741.0642	131.6287	130070.6	47665.56	-2.440
Y211D	825.2586	115.4343	158878.7	63039.08	+2.760
Y211T	806.8593	136.5648	167110.7	78849.29	+0.875
Y211A	654.1138	108.5343	108090.0	43661.09	+2.526
N348W/Y211A	594.9044	134.8233	108984.2	45451.00	+8.206

Before systematically applying SMD to the GP1/hTfR1 interaction, we needed to ensure the method was sufficiently sensitive to distinguish between relatively minor point mutations. While SMD has been applied previously to measure the binding energy of high-affinity T-cell receptor interactions [22,23], it is rarely used to parse small energy differences in a protein–protein interaction energy landscape. For this initial sensitivity analysis, we tested alanine substitutions congruent with the traditional experimental and computational approach.

We proceeded to compare our SMD results to that of the standard dual topology FEP approach to calculate relative free energy differences. The cor-

relation between the energetically rigorous FEP and our statistical approach is high. For all 11 complexes tested, the correlation between max force and FEP was $r = -0.795$ at $p = 0.0034$ (Figure 4.7), and the correlation between AUC and FEP was $r = -0.593$ at $p = 0.055$. Because of the strong correlation, we refer exclusively to the SMD results for the remainder of this work, focusing primarily on max force.

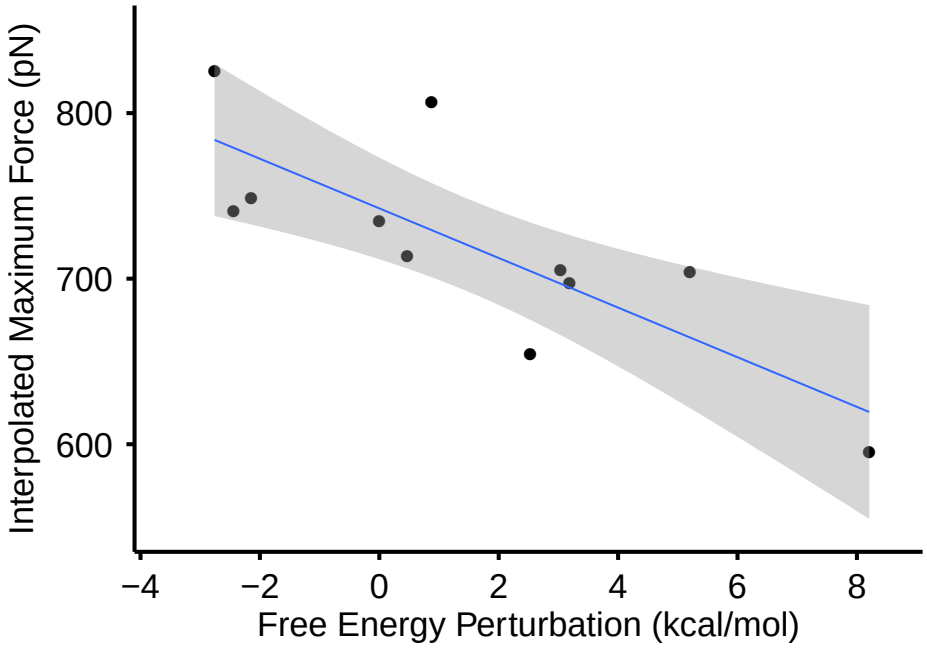

Figure 4.7: Max force versus free energy perturbation. Scatter plot of maximum force in SMD versus the relative free energy difference calculated by FEP for all 10 mutants tested plus the WT complex. The WT complex for FEP was simply set to 0.0. The correlation between the two is $r = -0.795$ with $p = 0.0034$.

We found that relative to WT, one alanine mutation (Y211A) produced

a very large and statistically significant difference in the maximum applied force and AUC (Figure 4.6, Table 4.3), while the other two did not (Table 4.3). When considering additional mutants (also discussed below), we found that maximum applied force was generally sufficient to distinguish mutants (Tables 4.3 and 4.4), and AUC was able to add a few more statistically significant differences (Table 4.5). In general, however, and consistent with the FEP results, maximum applied force seemed to be the more sensitive statistic than AUC.

Table 4.3: Pairwise differences (row variable minus column variable) in mean maximum applied force. Bolded values are statistically significant at $p < 0.05$.

	WT	N348A	N348W	N348K	vR111A	N348A/Y211A	vR111A/Y211A	Y211D	Y211T	Y211A
N348A	+14.036									
N348W	-29.414	-43.451								
N348K	-37.121	**-51.157**	-7.7060							
vR111A	-20.677	-34.713	+8.7370	+16.443						
N348A/Y211A	-30.782	-44.819	-1.3670	+6.3380	-10.105					
vR111A/Y211A	+6.5790	-7.4570	+35.993	+43.700	+27.256	+37.361				
Y211D	**+90.772**	**+76.736**	**+120.19**	**+127.89**	**+111.45**	**+121.56**	**+84.194**			
Y211T	**+72.373**	**+58.337**	**+101.79**	**+109.50**	**+93.051**	**+103.16**	**+65.795**	-18.399		
Y211A	**-80.371**	**-94.407**	-50.956	-43.250	**-59.694**	-49.588	**-86.950**	**-171.1 4**	**-152.75**	
N348W/Y211A	**-139.58**	**-153.62**	**-110.17**	**+102.46**	**-118.903**	**-108.80**	**+146.16**	**+230.35**	**-211.95**	**-59.209**

Table 4.4: Pairwise difference $p-$values for maximum applied force. Bolded values are statistically significant at $p < 0.05$.

	WT	N348A	N348W	N348K	vR111A	N348A/Y211A	vR111A/Y211A	Y211D	Y211T	Y211A
N348A	0.60									
N348W	0.31	0.077								
N348K	0.20	**0.038**	0.81							
vR111A	0.51	0.16	0.79	0.60						
N348A/Y211A	0.29	0.07	0.95	0.81	0.77					
vR111A/Y211A	0.82	0.79	0.21	0.13	0.35	0.20				
Y211D	**0.00093**	**0.0012**	**1.4x10^{-5}**	**5.0x10^{-6}**	**5.6x10^{-5}**	**1.2x10^{-5}**	**0.0022**			
Y211T	**0.01**	**0.018**	**0.00022**	**8.7x10^{-5}**	**0.0008**	**0.0002**	**0.021**	0.56		
Y211A	**0.0034**	**7.2x10^{-05}**	0.074	0.13	**0.035**	0.079	**0.0016**	**4.2x10^{-10}**	**4.2x10^{-8}**	
N348W/Y211A	**3.9x10^{-7}**	**1.1x10^{-10}**	**6.5x10^{-5}**	**0.00021**	**1.6x10^{-5}**	**7.2x10^{-5}**	**1.3x10^{-7}**	**¡ 2x10^{-16}**	**2.0x10^{-14}**	**0.036**

Table 4.5: Pairwise difference $p-$values for interpolated AUC. Bolded values are statistically significant at $p < 0.05$.

	WT	N348A	N348W	N348K	vR111A	N348A/Y211A	vR111A/Y211A	Y211D	Y211T	Y211A
N348A	0.33									
N348W	0.76	0.59								
N348K	0.59	0.80	0.76							
vR111A	0.55	0.85	0.76	0.94						
N348A/Y211A	**0.017**	0.07	**0.031**	0.076	0.08					
vR111A/Y211A	0.26	0.76	0.46	0.68	0.72	0.22				
Y211D	0.33	**0.029**	0.18	0.09	0.08	**0.00046**	**0.029**			
Y211T	0.09	**0.0056**	**0.046**	**0.027**	**0.023**	**4.1x10^{-5}**	**0.006**	0.59		
Y211A	**0.0056**	**0.027**	**0.016**	**0.029**	**0.031**	0.75	0.09	**8.2x10^{-5}**	**8.5x10^{-6}**	
N348W/Y211A	**0.006**	**0.029**	**0.017**	**0.032**	**0.034**	0.76	0.1	**9.4x10^{-5}**	**8.5x10^{-6}**	0.94

4.3.3 Comparative analysis of the GP1/hTfR1 interface

Considering the involvement of extended hydrogen-bonding networks in the GP1/hTfR1 interface (Figure 4.2), it was not clear that individual alanine mutations, even those that should destroy such networks, would significantly change the strength of interaction. One major advantage of first principles simulations is the ability to test mutations other than alanine without additional underlying assumptions in the energy function. As shown in Table 4.1, we made additional mutations based on biochemical intuition or available experimental data to chemically diverse amino acids including tryptophan, lysine, aspartate, and threonine. Several mutations caused significant relative affinity changes. In addition, to detect synergistic effects, we tested several double mutants where both mutations appeared to cause similar changes in binding. Then, we compared the size of those differences to single mutants (Figure 4.9 and 4.8).

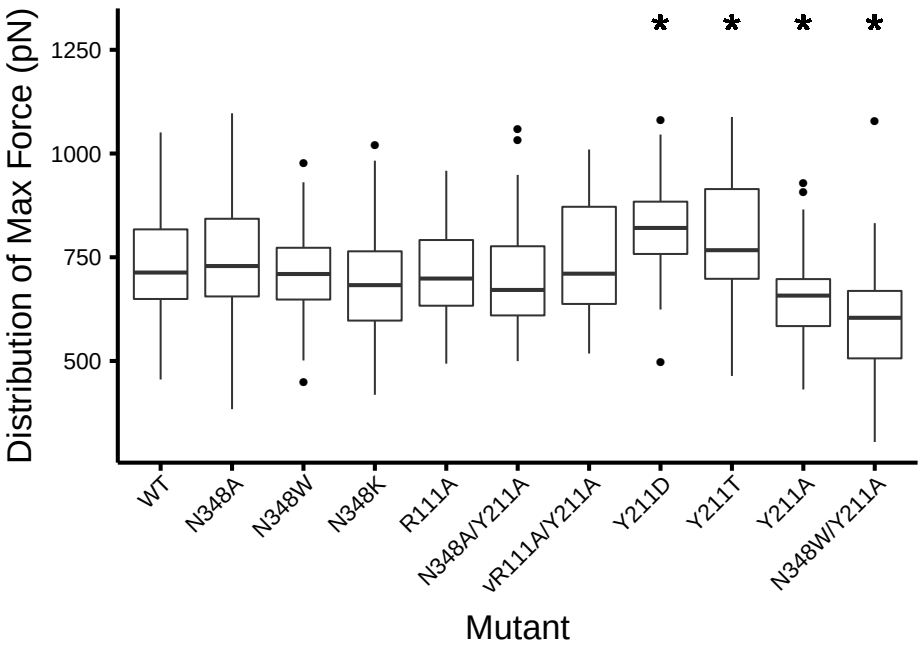

Figure 4.8: Distribution of interpolated maximum force for all bound complexes tested. Stars above the boxplots indicate a statistically significant difference in mean maximum force relative to the WT complex.

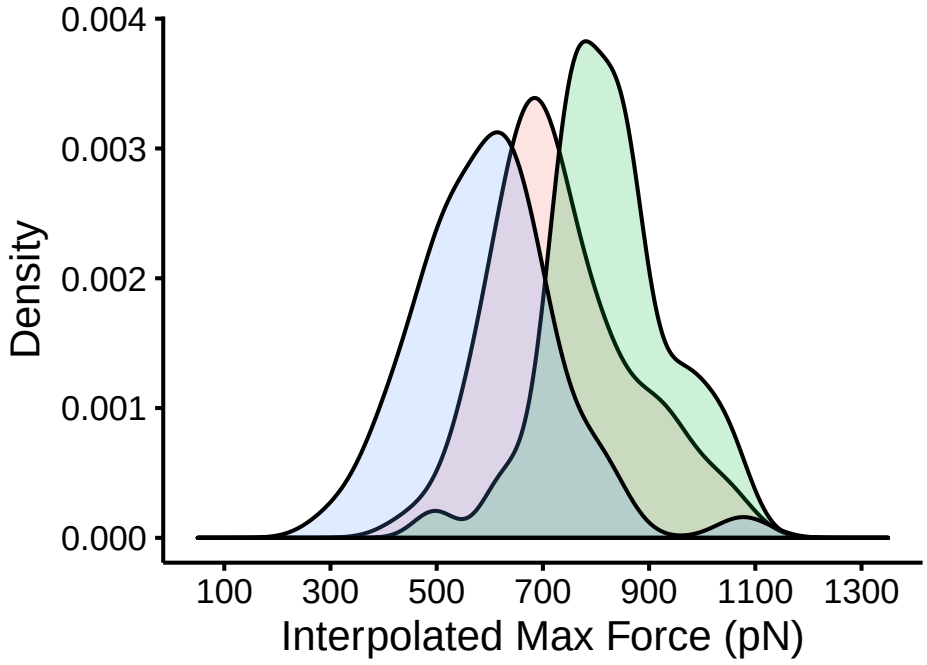

Figure 4.9: Distribution of interpolated maximum force for three different GP1/hTfR1 complexes. The WT GP1-hTfR1 complex in the middle is flanked by the tighter binding mutant Y211D on the right and the weaker binding double mutant N348W/Y211A on the left. The large non-overlapping areas indicate a large and statistically significant difference in these three complexes.

Although Y211A appears to have a large impact on binding affinity, no single mutant can provide enough evidence to understand the biochemical difference in binding mechanism. Since alanine is both smaller than tyrosine and also incapable of participating in hydrogen-bond interactions, we tested further mutations to identify the critical biochemical difference responsible for change in binding affinity. In particular, we substituted smaller side chains that, like tyrosine, were capable of hydrogen bonding. We chose Y211D and

Y211T, two mutations that have been discussed in the context of selection pressure on hosts in rodent populations [19, 78, 79]. Both mutations proved capable of causing a significant change in binding affinity in our simulations, but the change appeared to be increased affinity (Figures 4.9 and 4.8, and Table 4.4).

We also simulated several point mutations at N348 in the hTfR1. As discussed above, the alanine mutation at this site showed no significant difference in maximum applied force or AUC from WT (Tables 4.4 and 4.5). In addition, neither the N348Lys nor the N348W mutation showed a significant difference from WT. For both of these mutations, however, mean maximum applied force and mean AUC was lower than for WT (See Table 4.2). On the other hand, there was a detectable difference between N348A and N348Lys (Tables 4.4 and 4.5), with N348Lys being a weaker binder. Moreover, N348W showed nearly identical results to N348Lys. The mutations to large amino acids (N348W and N348Lys) produced nearly identical affinity changes, whereas the mutations to amino acids not capable of hydrogen bonding (N348A and N348W) produced significantly different affinity changes (Table 4.3). To check the consistency of our results, we hypothesized that the combination of Y211A and N348W, being chemically disconnected in two different hydrogen-bonding networks, would lead to a synergistic loss-of-binding. As expected, the double mutant was the weakest binding mutant tested ($p < 10^{-6}$, Tables 4.4 and 4.5) in this study. Further, according to maximum applied force (but not AUC), the combination of Y211A and N348W also showed signifi-

cantly weaker binding than Y211A by itself (Tables 4.4 and 4.5). We suspect that the effect of N348W alone is near the limit of detection using our method. A larger number of replicates would possibly have resolved affinity differences between N348W and WT or other mutants more consistently.

Last, we further analyzed a single mutation in GP1, vR111A. As mentioned previously, in our simulations this mutant showed no significant change in either maximum applied force or AUC (Tables 4.4 and 4.5), even though both quantities were, on average, lower than in WT (Table 4.2). This result was somewhat surprising, since Y211A, presumably disrupting the same hydrogen-bonding network as vR111A, displayed a significant reduction in affinity. To probe the interaction between position 111 in the GP1 and position 211 in the hTfR1 further, we also tested the double mutant vR111A/Y211A. This double mutant showed affinity indistinguishable from WT and significantly higher than Y211A alone (Table 4.3). This result shows that the two sites do indeed interact, and that replacing the hydrogen-bonding network at these sites with a hydrophobic interaction could lead to comparable binding affinity.

4.4 Discussion

We have applied a method utilizing steering forces in all-atom molecular dynamics simulations to evaluate the effects of mutations at the GP1/hTfR1 interface. We modeled mutations at several sites in the GP1/hTfR1 interface, and verified that our computational protocol was sensitive enough to

distinguish point mutants in hTfR1. Further, we identified two test statistics, maximum applied force and AUC, that can be used as proxies for binding affinity. Both of these statistics correlate well with FEP, but offer the simplicity of not requiring a large commitment to planning for the theoretical issues inherent to free energy methods. We systematically tested several point mutations to understand their contribution to the binding interaction. In the case of N348Lys, we have shown that the static structure provides little insight into why this mutation causes loss-of-infectivity *in vivo*. While N348 appears to be involved in a hydrogen-bonding network in the static structure, change in binding at that site may actually be caused by size and charge restriction. We also found that a negatively polar residue at site 211 in hTfR1 seem critical for a tight binding interaction. Any non-polar mutation at Y211 in hTfR1 is likely to completely halt viral entry and dramatically decrease the chances of MACV infection.

Traditionally SMD has been either applied to compute equilibrium free energies via a non-equilibrium approximation [33, 72, 73], used to estimate protein stability through unfolding [60], or used to calculate the absolute free energy of small molecule ligand binding [26]. Likewise, others have used SMD to understand the process of binding and unbinding at a resolution unmatched by experiment [23, 33]. Here, we have shown that SMD can provide insight into the *relative* strength of protein–protein interactions. Via SMD, one can separate mutations whose likely effect is altered binding affinity with simple statistics like maximum force of separation. Thus, SMD may open avenues

for subsequent experimental work in some situations where FEP may be prohibitively difficult.

Our findings rationalize several effects observed in both infectivity data and rodent populations [19,78]. First, we found that some substitutions at positions 211 and 348 did affect the strength of receptor binding. However, the computational data suggest that the reason and nature of the effects at these two sites are very different. At position 211, mutations to non-polar residues cause a large change in binding. This is congruent with what is known from viral entry data [19, 78]. By contrast, mutations at position 348 need only be small to maintain WT binding. The ability to hydrogen bond appears to be insignificant. This can be inferred from the fact that Y211A paired with large (W) and positively charged (Lys) substitutions at position 348 results in a larger than expected synergistic difference. That is, the double mutant Y211A/N348W caused a much larger decrease in binding than we expected from either mutation individually. Third, the GP1 mutation vR111A causes a loss-of-infection during *in vitro* infectivity assays [79], yet it was indistinguishable from the WT complex in our simulations. Although Y211A was the most disruptive single mutant we tested, vR111A in the GP1 was able to restore mean maximum applied force to WT levels (Table 4.2), and to levels significantly higher than observed for Y211A alone.

We would like to emphasize here that we cannot expect perfect agreement between our simulations and the available experimental data, but the correspondence to a well established free energy method bolsters our con-

clusions. While we have shown that our method can distinguish individual point mutations, we do not know the limit of detection with our method. First, it is possible that some mutants display measurable phenotypic effects in experiments yet appear identical in simulation. More extensive sampling or refinement of the simulation protocol could help to differentiate such mutants (see also next paragraph). Second, the SMD method is fundamentally limited by the accuracy of our starting structure. Third, the available experimental data for the GP1/hTfR1 system were generally obtained from entry assays or whole-cell binding assays rather than molecular binding assays. A mutant may cause a phenotypic difference in infectivity without generating a signal by our method. For example, entry could be lost in the experimental system because the protein is grossly or partially misfolded. An additional analytical step with circular dichroism or an analogous technique could clarify such large-scale folding differences. Further, since our simulations start with a bound structure, any changes that may dramatically affect the rate of association (different folds, trafficking issues, etc.) or relative orientation of the two proteins would be underestimated by our method.

There are a few additional challenges for investigating host-virus interactions via molecular dynamics simulation. As with any atomistic simulation, there is going to be a fairly large noise-to-signal ratio. To reduce noise, one could further customize each simulation, e.g. by determining the optimal pulling speed. Furthermore, larger amounts of computational resources will have a direct and powerful impact on the strength of any atomistic study [48].

Such resources could come in the form of increased compute time, improved code, or customized hardware for floating point operations [88]. With improved resources, we could investigate thousands of individual permutations in the GP1/hTfR1 binding interface. In addition, with additional compute time it would be possible to incorporate equilibrium sampling approaches [13] or use brute force equilibrium approaches [32] to improve resolution.

For future studies, although our approach offers the simplicity of not requiring prior knowledge about a system of interest (other than a bound model), at this point SMD may not the best approach for many relative affinity calculations. To ensure one's results are independent of the dissociation path one selects would require computing the work of separation for all likely paths. Such an approach eventually requires using the Jarzynski inequality [47] to establish a lower limit for binding energy and would quickly become computationally inefficient for evaluating a large number of mutations in most systems. However, considering the strong correlation between FEP and SMD in this system, it may not be important to ensure one's results are path independent for relative affinity calculations, as long as the same path is used for all complexes.

More importantly, with no *a priori* knowledge of the appropriate number of equilibration samples, the best duration of equilibration, the appropriate number of pulling runs, or the best pulling speed means the computational expense in our SMD protocol may not be commensurate with the information provided. For example, another all atom approach that makes calculations

via short simulations of spatially restrained complexes has proven capable of generating relatively accurate binding affinities with less compute time than is required from our steering strategy [39,40]. That being said, there is no reason to believe this SMD approach to mutagenic studies could not be optimized to reduce computational expense. Further analysis will be needed to understand the lower limits of resources required for accurate predictions.

Chapter 5

Conclusion

5.1 Discussion

There are a number of important points to come from this work. First, it is not only possible to incorporate measures of protein structure into models of molecular evolution, it can be a relatively simple task. Moreover, by incorporating relative solvent accessibility into the Goldman-Yang model, we can account for 5-10% more of the variation in the data compared to structure-naive models. We saw an improved fit via AIC and we were able to identify sites that were more conserved than we would have expected based on the site's RSA. In addition, our RSA-based method was able to identify several sites in hemagglutinin that are near the sialic acid binding region that evolve much faster than their RSA would have predicted.

Second, we applied our method to compare the evolutionary rate variation in hemagglutinin in different host species. We found that despite small numbers of sequences, random effect likelihood models tend to correctly recapitulate evolutionary rates even when fixed effect likelihood models perform poorly. In addition, we found that homologous sites in different host species had similar, though not identical, evolutionary rates. By making a direct

site-for-site comparison between two different host species, we were able to quantify the extent to which having a similar overall structure could account for the evolutionary rate variation among sites in the same protein. It appears structure could account for up to 24% to 36% of the evolutionary rate variation in hemagglutinin. Moreover, this number is probably near the lower limit for the extent to which structure constrains the evolutionary rate of proteins. Since hemagglutinin has a very strong adaptive pressure from one season to the next, purifying selection for biophysical stability is probably more relaxed than it is in most mammalian genes.

Third, to interrogate the effects of substitutions similar to those in the previous two chapters, we applied steered molecular dynamics simulations to pull apart a protein–protein complex. We found that with 50 replicates or less, we could differentiate 4 out of 10 mutants from the WT complex. Furthermore, we were able to see differences in these complexes by applying a force in only one direction and using the simple metric of maximum applied force. We were able to show that those differences make sense in light of some existing experimental data from viral entry assays. And finally, our results were in good agreement with one well established technique for calculating the relative free energy differences among various related structures.

5.2 Future Work

Next, we plan to extend our previous work to simulating full evolutionary trajectories. We will use a protein–protein interaction model system

that is similar to that of chapter 4. We want to understand the extent to which host tropism constraints affect the evolutionary process of viral binding proteins. We will simulate the evolution of the viral protein while leaving multiple potential host receptors static. In each test case, we will mutate amino acids in the protein–protein complex and use SMD or some other method to analyze the energy of interaction and internal stability. Then, we will use the metropolis criterion to either accept or reject the test mutation. The fitness landscape will be defined by a sigmoidal function where any mutation that improves binding will be accepted; any mutation that reduces binding below some threshold will have an exponentially declining probability of acceptance. With multiple possible homologous host receptors, we can multiply the probabilities of binding to the receptors to arrive at a final probability of acceptance per mutation tested. We expect that enforcing binding to three different host receptors will slow the adaptive changes that allow switching from one host to another. Likewise, relaxing the contraint of binding to a competent host will free the virus to adapt quickly to a new host. In addition, as the mutational process will be carried out in DNA, we can calculate the evolutionary rate as in chapters 2 and 3. We expect to see the evolutionary rate in the protein–protein interface will be even more constrained than residues with similar RSA in the core of the protein. Moreover, we expect that increasing the number of binding partners will more negatively constrain the evolutionary rate of interface sites. We hope simulations in a controlled evolutionary environment will inform our initial evolutionary rate studies; furthermore, they allow us to determine the

extent to which we should expect purifying selection on biophysical stability to depress to rate of adaptive evolution in viral proteins.

Bibliography

[1] J. Abraham, K. D. Corbett, M. Farzan, H. Choe, and S. C. Harrison. Structural basis for receptor recognition by new world hemorrhagic fever arenaviruses. *Nature Structural and Molecular Biology*, 17:438–444, 2010.

[2] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:6:716–723, 1974.

[3] J. M. Azais, E. Gassiat, and C. Mercadier. The likelihood ratio test for general mixture models with or without structural parameter. *ESAIM: Probability and Statistics*, 13:301–327, 2009.

[4] C. L. Bajaj, R. Chowdhury, and V. Siddahanavalli. F^2Dock: Fast fourier protein-protein docking. *IEEE Transactions on Computational Biology and Bioinformatics*, 8:45–58, 2011.

[5] A. Bakan, L. M. Meireles, and I. Bahar. ProDy: Protein dynamics inferred from theory and experiments. *Bioinformatics*, 27:1575–1577, 2011.

[6] L. Bao, H. Gu, K. A. Dunn, and J. P. Bielawski. Methods for selecting fixed-effect models for heterogeneous codon evolution, with comments on

their application to gene and genome data. *BMC Evolutionary Biology*, 7:S5, 2007.

[7] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57:289–300, 1995.

[8] S. Bhatt, E. C. Holmes, and O. G. Pybus. The genomic rate of molecular adaptation of the human influenza A virus. *Molecular Biology and Evolution*, 28:2443–2451, 2011.

[9] J. D. Bloom, D. A. Drummond, F. H. Arnold, and C. O. Wilke. Structural determinants of the rate of protein evolution in yeast. *Molecular Biology and Evolution*, 23:1751–1761, 2006.

[10] J. D. Bloom, L. I. Gong, and D. Baltimore. Permissive secondary mutations enable the evolution of influenza oseltamivir resistance. *Science*, 328:1272–1275, 2010.

[11] G. E. P. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society*, 26:211–252, 1964.

[12] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, 4:187–217, 1983.

[13] I. Buch, S. K. Sadiq, and G. De Fabritiis. Optimized potential of mean force calculations for standard binding free energies. *Journal of Chemical Theory and Computation*, 7:1765–1772, 2011.

[14] K. P. Burnham and D. R. Anderson. Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods and Research*, 33:261–304, 2004.

[15] R. M. Bush, C. A. Bender, K. Subbarao, N. J. Cox, and W. M. Fitch. Predicting the evolution of human influenza A. *Science*, 286:1921–1925, 1999.

[16] C. D. Bustamante, J. P. Townsend, and D. L. Hartl. Solvent accessibility and purifying selection within proteins of *Escherichia coli* and *Salmonella enterica*. *Molecular Biology and Evolution*, 17:301–308, 2000.

[17] R. N. Charrel and X. de Lamballerie. Arenaviruses other than Lassa virus. *Antiviral Research*, 57:89–100, 2003.

[18] J. D. Chodera, D. L. Mobley, M. R. Shirts, R. W. Dixon, K. Branson, and V. S. Pande. Alchemical free energy methods for drug discovery: progress and challenges. *Current Opinion in Structural Biology*, 21:150–160, 2011.

[19] H. Choe, S. Jemielity, J. Abraham, S. R. Radoshitsky, and M. Farzan. Transferrin receptor 1 in the zoonosis and pathogenesis of new world

hemorrhagic fever arenaviruses. *Current Opinion in Microbiology*, 12:467–482, 2011.

[20] S. C. Choi, A. Hobolth, D. M. Robinson, H. Kishino, and J. L. Thorne. Quantifying the impact of protein tertiary structure on molecular evolution. *Molecular Biology and Evolution*, 24:1769–1782, 2007.

[21] G. C. Conant and P. F. Stadler. Solvent exposure imparts similar selective pressures across a range of yeast proteins. *Molecular Biology and Evolution*, 26:1155–1161, 2009.

[22] M. Cuendet and O. Michielin. Protein-protein interaction investigated by steered molecular dynamics: The TCR-PMHC complex. *Biophysical Journal*, 95:3575–3590, 2008.

[23] M. Cuendet and V. Zoete. How T-cell receptors interact with peptide-MHCs: A multiple steered molecular dynamics study. *Proteins*, 79:3007–3024, 2011.

[24] A. M. Dean, C. Neuhauser, E. Grenier, and G. B. Golding. The pattern of amino acid replacements in α/β-barrels. *Molecular Biology and Evolution*, 19:1846–1864, 2002.

[25] W. Delport, K. Scheffler, and C. Seoighe. Models of coding sequence evolution. *Briefs in Bioinformatics*, 10:97–109, 2009.

[26] S. B. Dixit and C. Chipot. Can absolute free energies of association be estimated from molecular mechanical simulations? The biotin-

streptavidin system revisited. *Journal of Physical Chemistry A*, 105:9795–9799, 2001.

[27] S. Duffy, L. A. Shackelton, and E. C. Holmes. Rates of evolutionary change in viruses: patterns and determinants. *Nature Reviews Genetics*, 9:267–276, 2008.

[28] R. C. Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32:1792–1797, 2004.

[29] E. A. Franzosa and Y. Xia. Structural determinants of protein evolution are context-sensitive at the residue level. *Molecular Biology and Evolution*, 26:2387–2395, 2009.

[30] J. Gao, K. Kuczera, B. Tidor, and M. Karplus. Hidden thermodynamics of mutant proteins: A molecular dynamics analysis. *Science*, 244:1069–1072, 1989.

[31] M. K. Gilson, J. A. Given, B. L. Bush, and J. A. McCammon. The statistical-thermodynamic basis for computation of binding affinities: A critical review. *Biophysical Journal*, 72:11047–1069, 1997.

[32] T. Giorgino, I. Buch, and G. De Fabritiis. Visualizing the induced binding of SH2-phosphopeptide. *Journal of Chemical Theory and Computation*, 8:1171–1175, 2012.

[33] T. Giorgino and G. De Fabritiis. A high-throughput steered molecular dynamics study on the free energy profile of ion permeation through

Gramicidin A. *Journal of Chemical Theory and Computation*, 7:1943–1950, 2011.

[34] N. Goldman, J. L. Thorne, and D. T. Jones. Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics*, 149:445–458, 1998.

[35] N. Goldman and Z. Yang. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution*, 11:725–736, 1994.

[36] J. A. Grahnen, P. Nandakumar, J. Kubelka, and D. A. Liberles. Biophysical and structural considerations for protein sequence evolution. *BMC Evolutionary Biology*, 11:2455–2464, 2011.

[37] D. Gront, D. W. Kulp, R. M. Vernon, C. E. M. Strauss, and D. Baker. Generalized fragment picking in Rosetta: Design, protocols and applications. *PLOS ONE*, 6, 2011.

[38] J. Gumbart, E. Schreiner, D. N. Wilson, R. Beckmann, and K. Schulten. Mechanisms of SecM-mediated stalling in the ribosome. *Biophysical Journal*, 203:331–341, 2012.

[39] J. C. Gumbart, B. Roux, and C. Chipot. Efficient determination of protein-protein standard binding free energies from first principles. *Journal of Chemical Theory and Computation*, 9:3780–3798, 2013.

[40] J. C. Gumbart, B. Roux, and C. Chipot. Standard binding free energies from computer simulations: What is the best strategy? *Journal of Chemical Theory and Computation*, 9:794–802, 2013.

[41] S. E. Hensley, S. R. Das, A. L. Bailey, L. M. Schmidt, H. D. Hickman, A. Jayaraman, K. Viswanathan, R. Raman, R. Sasisekharan, J. R. Bennink, and J. W. Yewdell. Hemagglutinin receptor binding avidity drives influenza A virus antigenic drift. *Science*, 326:734–736, 2009.

[42] J. P. Huelsenbeck, S. Jain, S. W. D. Frost, and S. L. Kosakovsky Pond. A Dirichlet process model for detecting positive selection in protein-coding DNA sequences. *PNAS USA*, 103:6263–6268, 2006.

[43] W. Humphrey, A. Dalke, and K. Schulten. VMD - Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996.

[44] H. Hwang, T. Vreven, J. Janin, and Z. Weng. Protein-protein docking benchmark version 4.0. *Proteins*, 78:3111–3114, 2010.

[45] B. Isralewitz, J. Baudrya, J. Gullingsruda, D. Kosztinav, and K. Schulten. Steered molecular dynamics investigations of protein function. *Journal of Molecular Graphics*, 19:13–25, 2001.

[46] B. Isralewitz, M. Gao, and K. Schulten. Steered molecular dynamics and mechanical functions of proteins. *Current Opinion in Structural Biology*, 25:225–230, 2001.

[47] C. Jarzynski. Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach. *Physical Review E*, 56:5018–5035, 1997.

[48] M. O. Jensen, V. Jogini, D. W. Borhani, A. E. Leffler, R. O. Dror, and D. E. Shaw. Mechanism of voltage gating in potassium channels. *Science*, 336:229–233, 2012.

[49] W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577–2637, 1983.

[50] T. Kortemme, D. E. Kim, and D. Baker. Computational alanine scanning of protein-protein interfaces. *Science Signaling*, 2004:12, 2004.

[51] S. Kosakovsky Pond and S. D. Frost. Not so different after all: A comparison of methods for detecting amino acid sites under selection. *Molecular Biology and Evolution*, 22:1208–1222, 2005.

[52] S. Kosakovsky Pond and S. V. Muse. Site-to-site variation of synonymous substitution rates. *Molecular Biology and Evolution*, 22:2375–2385, 2005.

[53] S. L. Kosakovsky Pond, S. D. W. Frost, and S. V. Muse. HyPhy: hypothesis testing using phylogenetics. *Bioinformatics*, 21:676–679, 2005.

[54] S. L. Kosakovsky Pond, A. F. Y. Poon, A. J. L. Brown, S. D. W. Frost, and S. V. Muse. A maximum likelihood method for detecting directional

evolution in protein sequences and its application to influenza A virus. *Molecular Biology and Evolution*, 25:1809–1824, 2008.

[55] S. L. Kosakovsky Pond, K. Scheffler, M. B. Gravenor, A. F. Y. Poon, and S. D. W. Frost. Evolutionary fingerprinting of genes. *Molecular Biology and Evolution*, 27:520–536, 2010.

[56] S. Kryazhimskiy, G. A. Bazykin, J. Plotkin, and J. Dushoff. Directionality in the evolution of influenza A haemagglutinin. *Proceedings of the Royal Society B*, 275:2455–2464, 2008.

[57] S. Kryazhimskiy, J. Dushoff, G. A. Bazykin, and J. Plotkin. Prevalence of epistasis in the evolution of influenza a surface proteins. *PLOS Genetics*, 7:e1001301, 2011.

[58] W. G. Laver, G. M. Air, T. A. Dopheide, and C. W. Ward. Amino acid sequence changes in the haemagglutinin of A/Hong Kong (H3N2) influenza virus during the period of 1968-77. *Science*, 283:454–457, 1980.

[59] P. Liu, F. Dehez, W. Cai, and C. Chipot. A toolkit for the analysis of free-energy perturbation calculations. *Journal of Chemical Theory and Computation*, 8:2606–2616, 2012.

[60] H. Lu and K. Schulten. Steered molecular dynamics simulations of force-induced protein domain unfolding. *Proteins: Structure, Function, and Bioinformatics*, 35:453–463, 1999.

[61] N. Lu, D. A. Kofke, and T. B. Woolf. Improving the efficiency and reliability of free energy perturbation calculations using overlap sampling methods. *Journal of Computational Chemistry*, 25:28–39, 2004.

[62] M. Matrosovich, A. Tuzikov, N. Bovin, A. Gambaryan, A. Klimov, M. R. Castrucci, I. Donatelli, and Y. Kawaoka. Early alterations of the receptor-binding properties of H1, H2, and H3 avian influenza virus hemagglutinins after their introduction into mammals. *Journal of Virology*, 74:8502–8512, 2000.

[63] M. N. Matrosovich, A. S. Gambaryan, S. Teneberg, V. E. Piskarev, S. S. Yamnikova, D. K. Lvov, J. S. Robertson, and K.-A. Karlsson. Avian influenza A viruses differ from human viruses by recognition of sialyloligosaccharides and gangliosides and by a higher conservation of the HA receptor-binding site. *Virology*, 233:224–234, 1997.

[64] A. G. Meyer, E. T. Dawson, and C. O. Wilke. Cross-species comparison of site-specific evolutionary-rate variation in influenza hemagglutinin. *Philosophical Transactions of the Royal Society B*, 368:1614, 2013.

[65] A. G. Meyer, S. L. Sawyer, A. D. Ellington, and C. O. Wilke. Analyzing machupo virus-receptor binding by molecular dynamics simulations. *PeerJ*, 2:e266, 2014.

[66] A. G. Meyer and C. O. Wilke. Integrating sequence variation and protein structure to identify sites under selection. *Molecular Biology and Evolution*, 30:36–44, 2013.

[67] N. Michaud-Agrawal, E. J. Denning, T. B. Woolf, and O. Beckstein. MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. *Journal of Computational Chemistry*, 32:2319–2327, 2011.

[68] L. A. Mirny and E. I. Shakhnovich. Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *Journal of Molecular Evolution*, 291:177–196, 1999.

[69] S. V. Muse and B. S. Gaut. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution*, 11:715–724, 1994.

[70] R. Nielsen and Z. Yang. Likelihood models for detecting positive selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics*, 148:929–936, 1998.

[71] G. Mi no, M. Baez, and G. Gutierrez. Effect of mutation at the interface of Trp-repressor dimeric protein: a steered molecular dynamics simulation. *European Biophysics Journal*, 42:683–690, 2013.

[72] S. Park, F. Khalili-Araghi, E. Tajkhorshid, and K. Schulten. Free energy calculation from steered molecular dynamics simulations using Jarzynski's equality. *Journal Chemical Physics*, 119:3559–3567, 2003.

[73] S. Park and K. Schulten. Calculating potentials of mean force from steered molecular dynamics simulations. *Journal Chemical Physics*, 120:5946–5961, 2004.

[74] D. A. Pearlman. A comparison of alternative approaches to free energy calculations. *Journal of Physical Chemistry*, 98:1487–1493, 1989.

[75] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale, and K. Schulten. Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry*, 26:1781–1802, 2005.

[76] M. Porto, H. E. Roman, M. Vendruscolo, and U. Bastolla. Prediction of site-specific amino acid distributions and limits of divergent evolutionary changes in protein sequences. *Molecular Biology and Evolution*, 22:630–638, 2004.

[77] O. G. Pybus and A. Rambaut. Evolutionary analysis of the dynamics of viral infectious disease. *Nature Reviews Genetics*, 10:540–550, 2009.

[78] S. R. Radoshitsky, J. H. Kuhn, C. F. Spiropoulou, C. G. Albarino, D. P. Nguyen, J. Salazar-Bravo, T. Dorfman, A. S. Lee, E. Wang, S. R. Ross, H. Choe, and M. Farzan. Receptor determinants of zoonotic transmission of new world hemorrhagic fever arenaviruses. *Proceedings of the National Academy of Sciences*, 19:2664–2669, 2008.

[79] S. R. Radoshitsky, L. E. Longobardi, J. H. Kuhn, C. Retterer, J. C. Clester, J. Carra K. Kota, and S. Bavari. Machupo virus glycoprotein determinants for human transferrin receptor 1 binding and cell entry. *PLOS ONE*, 6, 2011.

[80] D. C. Ramsey, M. P. Scherrer, T. Zhou, and C. O. Wilke. The relationship between relative solvent accessibility and evolutionary rate in protein evolution. *Genetics*, 188:479–488, 2011.

[81] D. M. Robinson, D. T. Jones, H. Kishino, N. Goldman, and J. L. Thorne. Protein evolution with dependence among codons due to tertiary structure. *Molecular Biology and Evolution*, 20:1692–1704, 2003.

[82] N. Rodrigue, C. L. Kleinman, H. Philippe, and N. Lartillot. Computational methods for evaluating phylogenetic models of coding sequence evolution with dependence between codons. *Molecular Biology and Evolution*, 26:1663–1676, 2009.

[83] N. Rodrigue, N. Lartillot, D. Bryant, and H. Philippe. Site interdependence attributed to tertiary structure in amino acid sequence evolution. *Gene*, 347:207–217, 2005.

[84] N. Rodrigue, H. Philippe, and N. Lartillot. Assessing site-interdependent phylogenetic models of sequence evolution. *Molecular Biology and Evolution*, 23:1762–1775, 2006.

[85] N. Rodrigue, H. Philippe, and N. Lartillot. Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *Proceedings of the National Academy of Sciences*, 107:4629–4634, 2010.

[86] M. P. Scherrer, A. G. Meyer, and C. O. Wilke. Modeling coding-sequence evolution within the context of residue solvent accessibility. *BMC Evolutionary Biology*, 12:179, 2012.

[87] Schrödinger. The PyMOL molecular graphics system, version 1.3r1. Schrödinger, LLC., August 2010.

[88] D. E. Shaw, R. O. Dror, J. K. Salmon, J. P. Grossman, K. M. Mackenzie, J. A. Bank, C. Young, M. M. Deneroff, B. Batson, K. J. Bowers, E. Chow, M. P. Eastwood, D. J. Ierardi, J. L. Klepeis, J. S. Kuskin, R. H. Larson, K. Lindorff-Larsen, P. Maragakis, M. A. Moraes, S. Piana, Y. Shan, and B. Towles. Millisecond-scale molecular dynamics simulations on anton. *Proceedings of the Conference on High Performance Computing, Networking, Storage and Analysis*, SC09:39, 2011.

[89] T. Sikosek and H. S. Chan. Biophysics of protein evolution and evolutionary protein biophysics. *Journal of the Royal Society Interface*, 11, 2014.

[90] R. B. Squires, J. Noronha, V. Hunt, A. García-Sastre, C. Macken, N. Baumgarth, D. Suarez, B. E. Pickett, Y. Zhang, C. N. Larsen, A. Ramsey, L. Zhou, S. Zaremba, S. Kumar, J. Deitrich, E. Klem, and R. H.

Scheuermann. Influenza research database: an integrated bioinformatics resource for influenza research and surveillance. *Influenza and Other Respiratory Viruses*, 2012.

[91] A. Stamatakis. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22:2688–2690, 2006.

[92] J. Stevens, A. L. Corper, C. F. Basler, J. K. Taubenberger, P. Palese, and I. A. Wilson. Structure of the uncleaved human H1 hemagglutinin from the extinct 1918 influenza virus. *Science*, 303:1866–1870, 2004.

[93] B. J. Strait and T. G. Dewey. The Shannon information entropy of proteins sequences. *Biophysical Journal*, 71:148–155, 1996.

[94] Y. Suzuki. Natural selection on the influenza virus genome. *Molecular Biology and Evolution*, 23:1902–1911, 2006.

[95] Y. Suzuki and T. Gojobori. A method for detecting positive selection at single amino acid sites. *Molecular Biology and Evolution*, 16:1315–1328, 1999.

[96] A. U. Tamuri, M. dos Reis, A. J. Hay, and R. A. Goldstein. Identifying changes in selective constraints: Host shifts in influenza. *PLOS Computational Biology*, 5:e1000564, 2009.

[97] J. L. Thorne, N. Goldman, and D. T. Jones. Combining protein evolution and secondary structure. *Molecular Biology and Evolution*, 13:666–673, 1996.

[98] M. Z. Tien, A. G. Meyer, D. K. Sydykova, S. J. Spielman, and C. O. Wilke. Maximum allowed solvent accessibilities of residues in proteins. *PLOS ONE*, 8:e80635, 2013.

[99] A. Toth-Petroczy and D. S. Tawfik. Slow protein evolutionary rates are dictated by surface-core association. *Proceedings of the National Academy of Sciences*, 108:11151–11156, 2011.

[100] T. Vreven, H. Hwang, B. G. Pierce, and Z. Weng. Prediction of protein-protein binding free energies. *Protein Science*, 21:396–404, 2012.

[101] T. Vreven, H. Hwang, and Z. Weng. Integrating atom-based and residue-based scoring functions for protein-protein docking. *Protein Science*, 20:1576–1586, 2011.

[102] J. Wang, Y. Deng, and B. Roux. Absolute binding free energy calculations using molecular dynamics simulations with restraining potentials. *Biophysical Journal*, 91:2798–2814, 2006.

[103] J. R. R. Whittle, R. Zhang, S. Khurana, L. R. King, J. Manischewitz, H. Golding, P. R. Dormitzer, B. F. Haynes, E. B. Walter, M. A. Moody, T. B. Kepler, H. X. Liao, and S. C. Harrison. Broadly neutralizing human antibody that recognizes the receptor-binding pocket of influenza

virus hemagglutinin. *Proceedings of the National Academy of Sciences*, 108:14216–14221, 2011.

[104] H. Wickham. *ggplot2: Elegant graphics for data analysis*. Springer New York, 2009.

[105] D. C. Wiley, I. A. Wilson, and J. J. Skehel. Structural identification of antibody binding sites of Hong Kong influenza hemagglutinin and their involvement in antigenic variation. *Science*, 289:373–378, 1981.

[106] Z. Yang. Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. *Journal of Molecular Evolution*, 51:423–432, 2000.

[107] Z. Yang. *Computational Molecular Evolution*. Oxford University Press, 2006.

[108] Z. Yang and W. J. Swanson. Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Molecular Biology and Evolution*, 19:49–57, 2002.

[109] Z. H. Yang, R. Nielsen, N. Goldman, and A. M. K. Pedersen. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, 155:431–449, 2000.

[110] S. Zolla-Pazner and T. Cardozo. Structure–function relationships of HIV-1 envelope sequence-variable regions refocus vaccine design. *Nature Reviews Immunology*, 10:527–535, 2010.

Vita

Austin Garig Meyer graduated from Westlake High School in Austin, TX. He then attended Texas Tech University where he obtained a Bachelor of Science in Physics and a Bachelor of Arts in Philosophy in May of 2008. Next, he completed a Master of Science in Biotechnology in May of 2010 at Texas Tech University Health Sciences Center in Lubbock. The following summer he began graduate studies at the University of Texas at Austin first in Cell and Molecular Biology before moving to the Biochemistry program. In addition to graduate school, in 2012 he started medical school again at Texas Tech University Health Sciences Center. After completing his Ph.D., he expects to complete his medical degree in May of 2017.

Permanent address: austin.g.meyer@gmail.com

This dissertation was typeset with LaTeX[†] by the author.

[†]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.